



UNIVERSIDADE CATÓLICA PORTUGUESA

# Análise Preditiva na Empresa “OLI – Sistemas Sanitários, S.A.”

por

Teresa Sofia Henriques Ferreira da Rocha

Católica Porto Business School  
2020





UNIVERSIDADE CATÓLICA PORTUGUESA

# Análise Preditiva na Empresa “OLI – Sistemas Sanitários, S.A.”

Trabalho Final na modalidade de Relatório de Estágio  
apresentado à Universidade Católica Portuguesa  
para obtenção do grau de mestre em Gestão

por

Teresa Sofia Henriques Ferreira da Rocha

sob orientação de  
Maria da Conceição Andrade Silva

Universidade Católica Portuguesa, Católica Porto Business School  
julho e 2020



# Agradecimentos

À minha Orientadora e Professora Dr.<sup>a</sup> Maria Silva, pela sua disponibilidade e orientação excecional e exemplar. O seu profissionalismo e empatia foram os motores para a minha motivação e empenho.

Ao Dr. <sup>o</sup> Paulo Ribeiro, não só por me incentivar constantemente, mas também por me transmitir conhecimentos de várias áreas, fazendo-me apaixonar pelo Business Intelligence.

À Célia Fonseca, por me ajudar incessantemente e me proporcionar uma vivência empresarial bastante positiva.

À minha mãe, Maria Henriques. Qualquer descrição do que sinto por ela seria desatualizada, devido ao seu constante e exponencial crescimento. Ela é a bondade, a integridade, a tolerância, a harmonia e a terna melodia que apazigua o coração. Ser-lhe-ei eternamente grata.

Ao meu pai, ao meu exemplo, Filipe Rocha. É o melhor presente com o qual a vida me brindou. Dedico-lhe todas as minhas conquistas.

Ao Filipe Figueiredo, por me assoberbar de orgulho e ser como um irmão para mim.

À Ester Simões, por ser a minha melhor companheira e me preencher com a sua amizade e carinho insubstituíveis.

Ao Bruno Matos, ao Miguel Carvalho, ao Diogo Borges, por serem a minha segunda família.

À Mafalda Miranda, à Beatriz Carvalheira, ao João Queiroga e ao Miguel Pais pelo apoio incondicional demonstrado ao longo do meu percurso pessoal e académico.



# Resumo

A previsão constitui um importante ativo para as empresas, uma vez que ajuda no seu processo operacional e estratégico.

A previsão do futuro através de metodologias quantitativas caracteriza-se por ser bastante útil para a antecipação de tomadas de decisão, o que poderá levar a vantagem competitiva e sucesso no universo empresarial.

O desenvolvimento das áreas de *business intelligence* e *business analytics* têm um forte impacto na implementação destes métodos, uma vez que apresentam importantes ferramentas que se mostram eficientes para a análise e previsão.

Este trabalho centra-se em duas análises preditivas, uma referente às vendas totais da empresa “OLI – Sistemas Sanitários, S.A”, e a segunda referente à produção de energia do painel solar da “OLI Moldes, Lda”.

No sentido de averiguar métodos preditivos mais eficazes e eficientes para os dois casos de estudo implementados, empregou-se alguns métodos quantitativos de previsão existentes.

No primeiro caso, pode-se destacar a implementação do Método de Holt Winters Aditivo, o Método Holt Winters Multiplicativo e o Método ARIMA. No segundo caso foram utilizados métodos causais, como o Modelo de Modelo de Regressão Múltipla, o Modelo de Regressão de Ridge, o Modelo de Regressão de Lasso, o Modelo de Regressão de Elastic Net e o Modelo de Floresta Aleatória.

Deste modo, este trabalho conduz a uma elucidação dos conceitos de *business analytics* e *business intelligence*, que fortalecem a compreensão dos métodos quantitativos de previsão aplicados.

Palavras-chave: Business Analytics, Business Intelligence, Métodos Quantitativos de Previsão





# Abstract

Forecasting is an important asset for companies, as it helps in their operational and strategic process.

Predicting the future through quantitative methodologies is characterized by being very useful for anticipating the decision making, which can lead to competitive advantage and success in the business universe.

The development of the business analytics and business intelligence areas has a strong impact on the implementation of these methods, since they present important tools that prove to be efficient for analysis and forecasting.

This work focuses on two predictive analyses, one referring to the total sales of the company "OLI - Sistemas Sanitários, S.A", and the second one referring to the energy production of the solar panel of "OLI Moldes, Lda".

To ascertain more effective and efficient predictive methods for the two case studies implemented, some existing quantitative forecasting methods were used.

In the first case, we can highlight the implementation of the Additive Holt Winters Method, the Multiplicative Holt Winters Method and the ARIMA Method. In the second case, causal methods were used, such as the Multiple Regression Model, the Ridge Regression Model, the Lasso Regression Model, and the Random Forest Model.

In this way, this work leads to an elucidation of the concepts of business analytics and business intelligence, which strengthen the understanding of the quantitative forecasting methods applied.

**Keywords:** Business Analytics, Business Intelligence, Quantitative Forecasting Methods



# Índice

Agradecimentos .....	v
Resumo.....	vii
Abstract .....	ix
Índice .....	xi
Índice de Figuras.....	xiv
Índice de Tabelas .....	xvi
Introdução.....	18
1. Enquadramento Teórico .....	20
1.1 Introdução.....	20
1.2 Business Intelligence .....	21
1.3 Business Analytics .....	23
1.4 Análise Preditiva.....	27
1.5 R .....	27
2. Empresa “OLI – Sistemas Sanitários, S.A.” .....	30
2.1 História da Empresa.....	30
3. Métodos de Previsão .....	33
3.1 Introdução.....	33
3.2 Métodos Quantitativos .....	36
3.2.1 Séries Temporais .....	37
3.2.2 Método de Alisamento Exponencial .....	39
3.2.2.1 Método de Holt-Winters.....	40
3.2.3 Método Autorregressivo Integrado de Médias Móveis – ARIMA .....	42
3.2.3.1 Componente Autoregressiva (AR).....	42
3.2.3.2 Componente de Médias Móveis (MA) .....	44
3.2.3.3 Método Autorregressivo de Médias Móveis – ARMA .....	44
3.2.3.4 Método Autorregressivo Integrado de Médias Móveis – ARIMA..	45
3.2.3.5 Método Autorregressivo de Médias Móveis Sazonal - SARIMA....	46
3.2.3.6 Função de Autocorrelação.....	46
3.2.4 Modelos de Regressão .....	47
3.2.4.1 Modelo de Regressão Linear Simples.....	48
3.2.4.2 Modelo de Regressão Linear Múltipla .....	50
3.2.4.3 Modelo de Regressão de Lasso.....	52

3.2.4.4 Modelo de Regressão de Ridge .....	53
3.2.4.5 Modelo de Regressão da Floresta Aleatória .....	54
3.3 Métricas de Precisão .....	54
4. Aplicação do Caso de Estudo .....	59
4.1 Vendas da Empresa “OLI – Sistemas Sanitários, S.A.” .....	59
4.2 Consumos Energéticos da Empresa “OLI – Sistemas Sanitários, S.A.” .....	70
5. Conclusões .....	86
Bibliografia.....	88



# Índice de Figuras

Figura 1: Vetores de Business Analytics .....	26
Figura 2: Gráfico Temporal das Vendas Totais .....	60
Figura 3: Gráfico de Dados, Sazonalidade, Tendência e Componentes Irregulares .....	61
Figura 4: Subséries Temporais de cada Estação .....	62
Figura 5: Sazonalidade por Mês e Ano .....	63
Figura 6: Gráfico da Previsão do Método Holt Winters Aditivo .....	64
Figura 7: Gráfico da Previsão do Método Holt Winters Multiplicativo .....	66
Figura 8: Gráfico da Previsão do Método ARIMA.....	68
Figura 9: Gráfico de Comparação dos Métodos de Previsão.....	69
Figura 10: Dados do Consumo de Energia do Pannel Solar .....	71
Figura 11: Produção da Energia e as Informações Climatéricas .....	74
Figura 12: Correlações das Variáveis Climatéricas com a Produção de Energia .....	77



# Índice de Tabelas

Tabela 1: Previsão das Vendas Totais da Empresa, com o Método Holt Winters Aditivo.....	64
Tabela 2: Métricas de Precisão da Previsão das Vendas Totais da Empresa, com o Método Holt Winters Aditivo .....	66
Tabela 3: Previsão das Vendas Totais da Empresa, com o Método Holt Winters Multiplicativo.....	67
Tabela 4: Métricas de Precisão da Previsão das Vendas Totais da Empresa, com o Método Holt-Winters Multiplicativo e o Método Holt-Winters Aditivo	67
Tabela 5: Previsão das Vendas Totais da Empresa, com o Método ARIMA..	68
Tabela 6: Métricas de Precisão da Previsão das Vendas Totais da Empresa, com o Método ARIMA.....	68
Tabela 7: Comparação dos Métodos de Previsão.....	69
Tabela 8: Comparação de Métricas de Precisão.....	70
Tabela 9: Comparação do Resultados Preditivos dos Modelos Causais (Modelo de Regressão Linear Simples, Modelo de Regressão Linear Múltiplo, Modelo de Floresta Aleatória, Modelo de Regressão de Lasso e Modelo de Regressão de Ridge). .....	84
Tabela 10: Comparação das Métricas de Precisão dos Modelos Causais (Modelo de Regressão Linear Simples, Modelo de Regressão Linear Múltiplo, Modelo de Floresta Aleatória, Modelo de Regressão de Lasso e Modelo de Regressão de Ridge). .....	85





# Capítulo 1

## Introdução

Atualmente, a previsão desempenha um papel bastante importante nas atividades empresariais.

A previsão ajuda a preparar ações futuras por parte das empresas a todos os níveis, uma vez que prevê o acontecimento de eventos vindouros. Isto contribui para que as organizações estejam melhor preparadas e, é consequentemente, benéfico para o seu sucesso.

É com isto em mente que este trabalho concerne à previsão de vendas totais da empresa “OLI – Sistemas Sanitários, S.A”, uma empresa operacional há 66 anos no setor sanitário, e à previsão da produção de energia do painel solar da “OLI Moldes, Lda.”, uma sociedade limitada especialista na produção de moldes, fundada em 1991.

Para isso, foram levantadas várias questões, nomeadamente relativas à forma e quais as ferramentas a utilizar para realizar uma previsão para cada caso específico, e ao modo de determinação da fiabilidade dos métodos de previsão utilizados. Assim, este trabalho tem como finalidade fornecer um apoio instrumental para a realização de uma previsão adequada.

A estrutura do presente trabalho encontra-se dividida por capítulos.

O primeiro capítulo engloba as definições e caracterizações de conceitos primordiais para o entendimento do trabalho realizado, particularmente o

papel do *business analytics* e do *business intelligence*, a análise preditiva, e o programa utilizado para a consecução dos casos de estudo elaborados, o R.

O segundo capítulo apresenta o percurso da empresa até ao presente, para uma melhor perceção das suas operações.

O terceiro capítulo pretende elucidar relativamente às características da previsão, os métodos de previsão e métricas de precisão aplicadas.

O quarto capítulo concerne à aplicabilidade dos dois casos de estudo, mostrando os resultados obtidos dos métodos empregues.

O último capítulo corresponde às conclusões obtidas.

# Capítulo 1

## 1. Enquadramento Teórico

### 1.1 Introdução

O *business intelligence* e o *business analytics* (BI&A) são áreas que estão em constante crescimento e desenvolvimento, de extremo interesse para o contexto organizacional.

Chen et al. (2012) sugerem que as oportunidades associadas a dados e análises em diferentes organizações ajudaram a gerar um interesse significativo em BI&A.

O BI&A analisam dados críticos de negócios, em prol de ajudar uma empresa a entender melhor os seus negócios e mercados, e tomar decisões de negócios oportunas (Chen, H.L.Chiang, & C. Storey, 2012, pág. 1166).

Quando aliadas e funcionando de forma eficaz e eficiente, as ferramentas de BI&A, apoiam na planificação estratégica de toda a envolvente organizacional (Lee, 2013).

Lim et al. (2013), afirmam que BI&A consistem no desenvolvimento de tecnologias, sistemas, práticas e aplicações para analisar dados críticos de negócio, de modo a obter novos *insights* sobre negócios e mercados. De acordo com os mesmos autores, estes *insights* podem originar o aprimoramento de produtos e serviços, a eficiência operacional e fomentar relacionamentos com clientes.

Atualmente, o *business intelligence* pode ser considerado como um sistema de informação que possibilita ao *business analytics* atuar de forma eficiente.

Estes dois conceitos-chave identificam e organizam os dados, de forma a oferecerem soluções para a resolução de problemas internos da empresa. No entanto, têm pontos divergentes, como se irá ver nos pontos seguintes (1.2 e 1.3).

## 1.2 Business Intelligence

O conceito de *business intelligence* foi criado em 1865, por Richard Miller Devens, e sofreu grandes alterações ao longo do tempo.

Atualmente, é, por especialistas, considerado uma ampla categoria de aplicações e tecnologias para reunir, fornecer acesso e analisar dados, com o objetivo de ajudar os utilizadores corporativos a tomar melhores decisões de negócio (Ranjan, 2009).

Com efeito, o *business intelligence* poderá ser definido como uma agregação de conhecimento inteligente para a coleção, integração e análise dos dados, como indicam Olszak e Ziemba (2007). A exploração inteligente, integração, agregação e análise multidimensional de dados originários de vários recursos de informação são as suas funções fundamentais (ver M. Olszak & Ziemba, 2007, pág. 136).

Assim, é relevante abordar outro termo, o de *big data*. Este poder-se-á definir como o tamanho que está além da capacidade de ferramentas de *software* de banco de dados típicos para capturar, armazenar, gerir e analisar (Manyika, Chui, Brown, Bughin, & Dobbs, 2011).

Do mesmo modo, é necessário destacar o dinamismo deste termo que, aliado à tecnologia, vive um clima de constante mudança e evolução. Por isso, o

conceito de *big data* é considerado, por Hartmann et al. (2014), controverso, uma vez que a sua definição não é consensual.

No entanto, pode-se destacar três propriedades primordiais deste termo: o volume, a velocidade e a variedade.

O volume pode ser identificado como a quantidade de dados, enquanto que a velocidade permite o fluxo de informação processada.

Em alguns casos, o *big data* é definido pela capacidade de analisar uma variedade de conjuntos de dados, principalmente não estruturados, de fontes diversas, o que requer a capacidade de vincular conjuntos de dados e de extrair informações de dados que não possuem um modelo predefinido (explícito ou implícito). (Organization for Economic Co-operation and Development, 2013).

Os autores Rubin & Lukoianova (2013) reconhecem a importância da veracidade, aliada às propriedades do *big data*. A veracidade é uma construção teórica mais complexa, sem formas definidas de medi-la, especialmente, para grandes conjuntos de dados textuais não numéricos (Rubin & Lukoianova, 2013). De acordo com Rubin & Lukoianova (2013), é necessário definir as principais fontes de incerteza e os níveis de veracidade.

Segundo Ranjan (2009), o armazenamento de dados é a componente mais significativa de *business intelligence* que apoia a propagação física de dados através do tratamento dos numerosos registos de empresas para integração, limpeza, agregação e tarefas de consulta.

Ainda na ótica do *business intelligence*, surge o conceito de *data mart*. Segundo Ranjan (2009), o *data mart* contém dados operacionais que ajudam os especialistas de negócios a criar estratégias com base em análises de tendências e experiências históricas, baseada numa necessidade específica e predefinida.

Ainda de acordo com Ranjan (2009), as fontes de dados poderão ser bancos de dados operacionais, dados históricos, dados externos ou ser bancos de dados relacionais ou qualquer outra estrutura de dados que suporte a linha de

aplicações de negócios. Os dados também podem residir em muitas plataformas diferentes e podem conter informações estruturadas ou não estruturadas. (ver Ranjan, 2009, pág. 62).

Deste modo, o *business intelligence* agrupa diversas e abrangentes ferramentas úteis para a prossecução dos objetivos empresariais.

### 1.3 Business Analytics

De modo a complementar a análise e potenciar o seu entendimento, surge o *business analytics*.

Os sistemas de *business analytics* são uma estratégia de investimento importante para muitas organizações e podem potencialmente contribuir significativamente para o desempenho da empresa (Shanks, Cosic, & Maynard, 2012).

De acordo com Schläpke et al. (2013), o aumento da concorrência empresarial requer informações e análises de dados ainda mais rápidas e sofisticadas, as quais desafiam a gestão de desempenho a apoiar efetivamente o processo de tomada de decisão. *Business analytics* é um campo emergente que pode potencialmente estender o domínio da gestão de desempenho para fornecer uma melhor compreensão da dinâmica dos negócios, e levar a uma melhor tomada de decisão. (Schläpke, Silvi, & Möller, 2013)

Ainda seguindo a ótica destes autores, o *business analytics* é útil para prever a dinâmica que afeta os custos e que também pode ajudar a entender a dinâmica dos preços, bem como a relação já mencionada entre investimentos em marketing e retornos relacionados.

Numa vertente empresarial, os autores Schläpke et al. (2013) afirmam que os consideráveis desenvolvimentos de *business analytics* dos últimos anos,

fornececeram à gestão de desempenho, instrumentos promissores para lidar com os desafios atuais.

Assim, se estas abordagens forem incluídas nas pequenas e médias empresas, as mesmas fornecem novas percepções sobre a dinâmica dos negócios e o seu desempenho relacionado, que podem ser explorados e aproveitados, o que, por sua vez, pode resultar numa maior eficácia de gestão (Schláfke, Silvi, & Möller, 2013, pág. 111).

Assim, poder-se-á concluir que, sendo um processo inovador, o *business analytics* pode conceder à empresa vantagem competitiva e crescimento sustentável.

O *business analytics* também se destaca noutras realidades, sem ser a nível empresarial. Por exemplo, em instituições de educação e de saúde, pode ajudar todos os sistemas a serem cada vez mais aprimorados e automáticos.

De acordo com Yang e Moody (1999), a visualização de dados é muito importante para o ser humano entender as relações estruturais entre variáveis num sistema e eliminar alguns modelos irrealistas.

Portanto, a visualização de dados permite ao usuário obter informações de dados e chegar a novas hipóteses. (Keim, 2002)

A visualização de dados visa a clarificação de informação para qualquer propósito, a qualquer departamento e cargo.

Uma ferramenta basilar de *business analytics* são as *dashboards*. Segundo Few (2005), as *dashboards* fornecem um meio distinto e poderoso para comunicar informações, com benefícios específicos, mas também representam um conjunto específico de desafios de *design*.

Assim, é uma ferramenta visual que permite a exibição da informação necessária e organizada para que, tanto um administrador, como um gestor, consiga observá-la e entendê-la.



Desta forma, o objetivo é ter as informações mais importantes prontamente, sem esforço e imediatamente disponíveis e precisa de apontar rapidamente que algo merece atenção, que algo pode exigir ação (ver Few, 2005, pág. 20). Por exemplo, numa empresa em que tenha uma alta taxa de turnover, é necessário que este indicador se encontre na *dashboard*, para que se resolva a questão, de modo a melhorar a sua *performance*.

Uma outra ferramenta de *business analytics* é o *data mining*. De acordo com Lee (2013), o *data mining* deteta os padrões e relacionamentos num ou em vários conjuntos de dados volumosos. Segundo o mesmo autor, o *data mining* é um processo que utiliza técnicas de análise preditiva, um conceito descrito mais detalhadamente na secção 1.4, como a estatística, para apoiar nas tomadas de decisão e operações da empresa. Exemplificando, ao reconhecer movimentos padronizados dos clientes, permite verificar a adequação dos produtos/serviços para cada cliente, ajudando a empresa a indicar soluções mais apropriadas para cada destinatário. (Lee, 2013)

Além disso, algumas das técnicas tradicionais de data mining incluem classificação, agrupamento, análise de *outliers*, padrões sequenciais, análise de séries temporais, previsão, regressão, análise associações e métodos multidimensionais, incluindo o processamento analítico online (OLAP). (Ranjan, 2009).

Segundo Schläpke et al. (2013), os consideráveis desenvolvimentos de *business analytics* dos últimos anos, forneceram à gestão de desempenho, instrumentos promissores para lidar com os desafios atuais, como matemática, estatística, econometria, tecnologias de informação e ferramentas para a recolha e análise de dados.

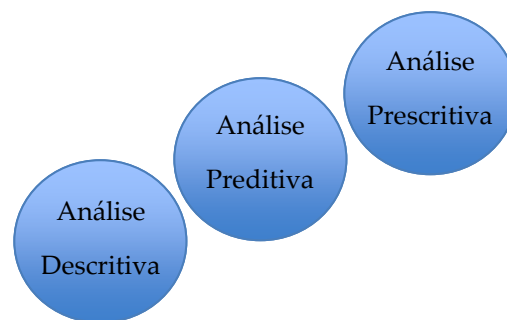
A inserção destes instrumentos nas pequenas e médias empresas fornece novas perceções sobre a dinâmica dos negócios e sobre o seu desempenho, que

podem ser explorados de forma a resultar numa maior eficácia de gestão (ver Schläfke, Silvi, & Möller, 2013, pág.111).

Os três vetores de *business analytics* são a análise descritiva, a análise preditiva e a análise prescritiva (figura 1).

Figura 1

Vetores de Business Analytics



A análise preditiva não só se centra na comunicação de eventos passados e do presente, mas também na categorização, caracterização, consolidação e classificação dos dados, através de ferramentas como as dashboards e relatórios (Lustig, Dietrich, Johnson & Dziekan, 2010).

Por outro lado, e de acordo com Lustig et al. (2010), o processo de foco da análise preditiva é a previsão, através dos acontecimentos passados, isto é, concentra-se no que irá acontecer no futuro. A agregação (*clustering*, em inglês), as árvores de decisão e as redes neurais são alguns mecanismos da análise preditiva (Lustig, Dietrich, Johnson & Dziekan, 2010).

Como Lustig et al. (2010) afirmam, a análise prescritiva, através do planeamento baseado na otimização, responde à forma como se deverá atuar no futuro, consoante as análises descritiva e preditiva. Esta análise, através de algoritmos matemáticos robustos, identifica um sistema de potenciais decisões, analisa as interações entre essas decisões, e determina as restrições de combinações dessas decisões, com o objetivo de procurar o melhor conjunto de

decisões num determinado contexto (Lustig, Dietrich, Johnson & Dziekan, 2010).

## 1.4 Análise Preditiva

Um importante conceito de reflexão que está aliada a outros termos aqui expostos é a análise preditiva. Como sugere Waller e Fawcett (2013), a análise preditiva é um subconjunto da ciência de dados, que ilumina algumas necessidades interessantes da pesquisa.

Este conceito pode gerar um retorno sobre o investimento substancial, já que pode ajudar as empresas a otimizar processos existentes, entender melhor o comportamento do cliente, identificar oportunidades inesperadas, e antecipar possíveis problemas (Boonsiritomachai, McGrath, & Burgess, 2016, pág. 4).

A análise preditiva funciona de forma indutiva, isto é, não presume nada sobre os dados. Emprega análises estatísticas, *machine learning*, computação neural, robótica, matemática computacional e técnicas de inteligência artificial para explorar todos os dados, em vez de um subconjunto restrito, para descobrir relações e padrões significativos. (ver Boonsiritomachai, McGrath, & Burgess, 2016, pág. 6).

## 1.5 R

De acordo com Slonneger e Kurtz (1995), as linguagens de programação podem definidas como linguagens artificiais, inicialmente com o objetivo de se comunicar com computadores, mas, igualmente importante, para comunicar algoritmos entre pessoas.

Os constituintes duma linguagem de programação são a sua sintaxe e a sua semântica.

A sintaxe define as relações formais entre os constituintes de uma linguagem, fornecendo, assim, uma descrição estrutural das várias expressões que compõem as cadeias legais na linguagem. (ver Slonneger & Kurtz, 1995, pág. 1). Deste modo, os mesmos autores explicam que a mesma lida apenas com a forma e estrutura dos símbolos num idioma, sem nenhuma consideração dada ao seu significado.

Ihaka e Gentleman (1996) realçam que a sintaxe de uma linguagem é inteiramente superficial, isto é, não concebe à sintaxe a aceção dos símbolos, no entanto, é importante pois determina a maneira como os utilizadores do idioma se expressam.

Por outro lado, a semântica revela o significado de cadeias sintaticamente válidas num idioma, descrevendo o comportamento que um computador segue ao executar um programa na linguagem. (ver Slonneger & Kurtz, 1995, pág. 1).

Apesar da linguagem de programação ser retratada pela sintaxe e pela semântica, ter boa estrutura sintática e semântica por si só não produzirá uma linguagem útil nem sua ausência criará uma linguagem inútil. (ver Ihaka & Gentleman, 1996, pág. 300). Assim, estes autores alegam a presença de demais elementos de relevante importância, nomeadamente as ferramentas, uma vez que propicia o processamento de cálculos.

A harmonia/agregação destas três componentes é designada de *run-time environment*, que é a visão típica do utilizador do programa. (Ihaka & Gentleman, 1996, pág. 300).

Sob outra perspetiva, o Slonneger e Kurtz (1995) referem que a pragmatística faz alusão aos aspetos da linguagem que envolvem os usuários da linguagem e inclui questões como facilidade de implementação, eficiência no aplicativo e metodologia de programação, integrando-se assim na definição da linguagem de programação.

Em 1991, Ross Ihaka e Robert Gentleman conceberam uma linguagem de programação – a linguagem R –, no Departamento de Estatística da Universidade de Auckland.

A linguagem R, como os criadores afirmam, é fortemente influenciado por duas línguas existentes – S, de Becker, Chambers e Wilks (1995) e Scheme, de Steel and Sussman (1975) (Ihaka & Gentleman, 1996).

Por um lado, a linguagem resultante é muito semelhante em relação a S, por outro, a implementação e a semântica subjacentes são derivadas do Scheme. (ver Ihaka & Gentleman, 1996, pág. 299).

Deste modo, a linguagem passou por um processo de metamorfose linguística, em que numa primeira fase, foi escrita e interpretada com base no Scheme e, postumamente, adaptada para assimilar-se com a linguagem S.

Assim, o R é uma linguagem e um programa com um vasto leque de bibliotecas, que se iniciou por ajudar em soluções para fins estatísticos e resoluções gráficas, tendo evoluído para opções mais abrangentes.

O ambiente em que é executado o R é intitulado de RStudio.

## Capítulo 2

### 2. Empresa “OLI – Sistemas Sanitários, S.A.”

#### 2.1 História da Empresa

A “OLI – Sistemas Sanitários, S.A.”, fundada a 1 de março de 1954, é uma empresa especializada em produtos sanitários, presente em 80 países, em 5 continentes.

A empresa está sediada em Aveiro, com cerca de 500 trabalhadores, e conta com três filiais: a “OLI Itália”, a “OLI Rússia” e a “OLI Alemanha”.

No início da década de 90, foi criada a empresa “OLI Moldes, Lda.”, dedicada à produção de moldes para a injeção de componentes em materiais termoplásticos. Estes moldes são utilizados para o fabrico dos materiais da “OLI-Sistemas Sanitários, S.A.”, os quais se distinguem pelo seu rigor e excelência no mercado.

Ainda, a “OLI Moldes, Lda.” destaca-se em serviços como a engenharia de produto, protótipos e estudos reológicos, inserindo-se em vários setores de negócio, tais como o automóvel, motociclos, eletrodomésticos, eletrónica e telecomunicações, construção e hidro-sanitário.

Em 1993, a “OLI – Sistemas Sanitários, S.A.” integrou-se no Grupo SILMAR, uma empresa sediada em Itália e focalizada em quatro esferas de atividade:

aquecimento, fundição em alumínio, metalização em plásticos e redes de esgotos e águas.

Sendo a maior produtora de autoclismos do sul da Europa, os seus produtos primordiais são os autoclismos interiores, os autoclismos exteriores e os autoclismos plásticos para tanques cerâmicos, as placas de comando, os módulos sanitários e os mecanismos. Oferece também uma ampla gama de soluções de banho, de equipamentos de climatização e de tubagens.

Distingue-se pela sua constante inovação e sustentabilidade, traçadas, nomeadamente, pelo desenvolvimento da dupla descarga de água do autoclismo, que diminui 50% de água consumida, e pela torneira de boia, que permite poupar até 9 litros de água por dia e 2% da faturação mensal.

Concentra-se sempre em desenvolver produtos de preocupação global, tais como o “Azor Plus”, um suporte de autoclismo que permite a adequação de medida, através dum comando.

Pela implementação de Kaizen em 2007, uma metodologia de melhoria contínua de origem japonesa, a empresa foi premiada pela categoria “Excelência na Produtividade”. Além disso, a empresa foi galardoada com 7 prémios relativos ao seu design inovador.

É uma das principais produtoras de equipamentos de descarga para a indústria cerâmica, o segundo fabricante europeu de autoclismos e o único fabricante português de autoclismos interiores. Isto traduz-se, em média e anualmente, quase 3 milhões de mecanismos, 2 milhões de autoclismos e mais de 310 000 unidades de autoclismos interiores.

É uma das empresas portuguesas com um maior registo de patentes, contando com 48, até ao momento. A sua contemporaneidade, o permanente desenvolvimento e investigação permite o seu crescimento consecutivo.

A formulação da sua estratégia passa pelo seu crescimento sustentável do volume de atividade e da rentabilidade, de forma equilibrada e reforçando a

estrutura financeira, melhorar a proposta de valor e aprimorar os níveis de produção, de inovação, de fiabilidade e de potência dos sistemas de informação e de contributo do *procurement* para a competitividade. A estratégia também recai no reforço do envolvimento e cooperação por parte dos elementos da OLI e a consolidação do processo de internacionalização das 3 filias e em Espanha.

Em 2019, a faturação da OLI foi de cerca de 59 milhões de euros. O seu rendimento de 2018 comparativamente a 2017, desceu 2%, aproximadamente mais de 1 milhão e 300 mil, o que resultou num decréscimo de 27% em termos de resultado líquido.



# Capítulo 3

## 3. Métodos de Previsão

### 3.1 Introdução

Atualmente, a previsão constituiu um ativo indispensável para o sucesso organizacional.

Dentro das organizações, uma questão importante é a previsão de vendas. (Caiado, 2016). Ainda, o autor acrescenta que esta previsão apoia a planear as ações operacionais e a compor instrumentos da gestão, como os orçamentos.

A previsão poderá incidir sob um intervalo de tempo curto, médio ou longo, conforme a questão em estudo e o objetivo a alcançar. Como os autores Hyndman e Athanasopoulos (2018) sugerem, as previsões de curto prazo utilizam-se para a gestão de pessoas tendo em conta as respetivas previsões de procura, e ainda a gestão de produção e de transporte.

Já as previsões de intervalo de tempo intermédio são relevantes para antecipar os bens e recursos humanos necessários. (Hyndman & Athanasopoulos, 2018).

Finalmente, as previsões de longo prazo têm uma maior amplitude, no sentido que englobam as ações estratégicas a tomar por parte da empresa,

relacionadas com a envolvente externa e interna. (Hyndman & Athanasopoulos, 2018)

Para a análise preditiva ser efetuada de forma correta, é necessário desenvolverem-se procedimentos e sistemas no decorrer do cumprimento de certas etapas (Hyndman & Athanasopoulos, 2018).

Assim, de acordo com os mesmos autores, as etapas da previsão são a definição do problema, a recolha de dados, a análise exploratória, a explicação/modelação dos modelos de previsão e a utilização e avaliação do modelo preditivo.

Para efetuar uma previsão, primeiramente, é crucial definir o problema em questão e, para isso e segundo Makridakis, Hyndman e Wheelwright (1998), é crucial conhecer a importância e a utilidade que a mesma tem para a organização, tendo em conta o destinatário que a sugeriu.

Após a definição do problema, deverá ser realizada a recolha de informações essenciais referentes à questão principal. As informações são, no mínimo, de dois tipos: dados estatísticos (geralmente numéricos) ou julgamentos e experiência acumulados dos especialistas (Makridakis, Hyndman & Wheelwright, 1998, pág. 13). O levantamento de todas as informações de relevância é crucial para a definição dos modelos matemáticos e estatísticos pertinentes (ver Caiado, 2016, pág.38).

Com os elementos informativos, deve-se proceder, primeiramente, a uma visualização gráfica de modo a termos uma visão holística dos dados a examinar, permitindo a observação de possíveis padrões, de modo imediato. (Hyndman & Athanasopoulos, 2018).

De acordo com os mesmos autores, seguidamente, é relevante uma síntese numérica dos dados, tendo como objetivo chegar ao entendimento dos dados que iremos tratar e comunicar as relações entre os dados. Esta constituiu o terceiro passo da previsão designada de análise preliminar ou exploratória.

A fase posterior é a escolha e a modelação dos modelos de previsão quantitativos. Segundo Hyndman e Athanasopoulos (2018), cada modelo é uma construção artificial que se baseia num conjunto de suposições (explícitas e implícitas) e geralmente envolve um ou mais parâmetros que devem ser estimados, utilizando os dados históricos conhecidos.

Por fim, a última fase é a utilização e avaliação do modelo preditivo. O modelo escolhido é testado de acordo com os seus resultados previstos, com base no comportamento anterior. Assim, em conjunto com o supervisionamento do modelo, é imperioso a adoção de métricas de precisão para avaliar o método utilizado, que irão ser abordadas singularmente no capítulo 3.3.

Conceitos indissociáveis à análise preditiva são os métodos preditivos, que podem variar em custo, complexidade e valor (Hyndman & Athanasopoulos, 2018). Há dois tipos de métodos de previsão: métodos qualitativos de previsão e métodos quantitativos de previsão.

Assim, a previsão poderá ser realizada via abordagens qualitativas, como a previsão de peritos (mercados preditivos e método delphi), analogias estruturadas e decomposição de julgamento, e via métodos quantitativos, como analogias quantitativas, modelos causais, e modelos de decomposição.

Os métodos qualitativos provêm de opiniões, convicções e do entendimento fundamentado de especialistas. (Makridakis, Wheelwright & Hyndman, 1998, pág. 12).

Estes métodos poderão ser adequados quando não existem dados quantitativos ou quando há uma mudança relevante do contexto do passado. Ainda são combinados entre si, ou complementados com as suas contrapartes quantitativas, para viabilizar a consecução de objetivos, em várias áreas (Makridakis, Wheelwright & Hyndman, 1998, pág. 12).

Os métodos quantitativos serão aprofundados ao longo deste capítulo.

## 3.2 Métodos Quantitativos

Neste trabalho, foram aplicados os métodos quantitativos de previsão, especialmente os métodos de Holt-Winters e ARIMA. Assim, os métodos quantitativos empregues são descritos com maior minúcia nos pontos seguintes.

De acordo com Makridakis, Wheelwright e Hyndman (1998), os métodos quantitativos de previsão devem ser aplicados quando existe informação numérica do passado e que seja razoável assumir o pressuposto da continuidade, isto é, que alguns aspetos dos dados passados se reproduzam no futuro.

Os métodos quantitativos dividem-se em métodos de séries temporais ou extrapolativos e em métodos causais ou explicativos.

De acordo com Caiado (2016), os métodos de séries temporais, como se irá verificar no ponto 3.2.1, utilizam dados quantitativos históricos e com o pressuposto de que os movimentos anteriores se mantêm posteriormente.

Já os modelos causais baseiam-se em variáveis explicativas, isto é, variáveis relacionadas e que expliquem a variável de interesse a ser estimada, como por exemplo, os modelos de regressão (ver Caiado, 2016).

Segundo Hyndman e Athanasopoulos (2018), a utilização dos métodos de séries temporais poderá ter vantagens sob os métodos causais. Nos métodos causais ou explicativos, a falácia “cum hoc ergo propter hoc”, que significa “com isto, logo por causa disto”, pode ser incorrida. Ainda, a adversidade de ter de se prever tanto as variáveis explicativas, como as variáveis de interesse são as razões pelas quais se poderá justificar a escolha do método de séries temporais, em deterioração dos métodos causais (Hyndman & Athanasopoulos, 2018).

### 3.2.1 Séries Temporais

Ao contrário dos métodos causais, os métodos de séries temporais centram-se nas informações quantitativas disponíveis num intervalo de tempo de uma variável de interesse.

Os modelos de alisamento exponencial e o ARIMA são exemplos de modelos de previsão realizados com séries temporais.

Segundo Chatfield (2000), uma série temporal é constituída por quatro componentes: a variação sazonal, a tendência, o movimento cíclico e as flutuações irregulares. São definidas, pelo autor, da seguinte maneira:

A sazonalidade observa-se quando, no nosso conjunto de observações, há comportamentos que se repetem em períodos homólogos. Assim, é necessário que se observe os movimentos durante vários anos, para se verificar esta variação. Por exemplo, entre os anos 2002 até 2019, na empresa em análise, verifica-se que há um decréscimo constante no mês de agosto.

A tendência existe quando há um crescimento ou um decréscimo sucessivo dos dados, num longo período.

A variação cíclica verifica-se quando movimentos de ciclos em períodos diferentes de um ano estão presentes. Nesta componente estão enquadradas épocas de expansão e recessão económica.

As flutuações irregulares são observações anómalas que não se registam dentro da tendência, da sazonalidade, e de outros efeitos sistemáticos.

A sazonalidade repete-se de forma constante num certo momento do tempo consecutivamente, enquanto que com as variações cíclicas tal não se verifica. A amplitude a extensão temporal dos ciclos tende a ser maior que os movimentos sazonais (ver Hyndman & Athanasopoulos, 2018).

De outra perspetiva, os autores Hyndman e Athanasopoulos (2018), afirmam que, ao decompor uma série temporal em componentes, se verifica a

combinação das componentes tendência e ciclo. Assim, os constituintes duma série temporal são as componentes “tendência-ciclo”, também denominada de tendência apenas, a sazonalidade e as flutuações irregulares, também designada de observações anómalas ou componente remanescente (Hyndman e Athanasopoulos, 2018).

A decomposição de séries temporais é um processo que decompõe as componentes anteriormente abordadas das séries temporais, de forma aditiva ou multiplicativa.

A decomposição aditiva pode ser descrita em (1), e a decomposição multiplicativa em (2), onde  $Y_t$  representa os valores observados,  $S_t$  a componente sazonal,  $T_t$  a componente do tendência-ciclo e a  $R_t$  componente remanescente, no período  $t$ .

$$Y_t = S_t + T_t + R_t$$

(1)

$$Y_t = S_t \times T_t \times R_t$$

(2)

Segundo Hyndman e Athanasopoulos (2018), a decomposição aditiva é mais apropriada se a magnitude das flutuações sazonais ou a variação em torno da tendência-ciclo não variar com o nível da série temporal. Assim, os autores indicam que quando a variação no padrão sazonal, ou a variação em torno da tendência-ciclo, parece ser proporcional ao nível da série temporal, uma decomposição multiplicativa é mais apropriada.

Ainda de acordo com Hyndman e Athanasopoulos (2018), uma alternativa ao uso de uma decomposição multiplicativa é primeiro transformar os dados até que a variação da série pareça estável ao longo do tempo, e depois usar uma decomposição aditiva. Quando uma transformação de *log* é usada, os mesmos

autores afirmam que equivale a usar uma decomposição multiplicativa, uma vez que,

$$\log Y_t = \log S_t + \log T_t \times \log R_t$$

(3)

Segundo descreve Chatfield (2020), os objetivos das séries temporais são a caracterização dos dados através de estatísticas sumárias e/ou gráficos, a escolha de um modelo estatístico apropriado para o processo de geração de dados, a previsão de eventos futuros e o controlo da eficácia da previsão.

### 3.2.2 Método de Alisamento Exponencial

Em 1944, os métodos de alisamento exponencial foram criados por Robert G. Brown, enquanto trabalhava na marinha dos Estados Unidos. Posteriormente, na década de 50, incluiu as propriedades de tendência e sazonalidade, e estendeu as séries temporais contínuas a discretas (Hyndman, Koehler, Ord, & Snyder, 2008).

De acordo com Hyndman et al. (2008), Charles Holt, professor da Universidade de Texas, em conjunto com Peter Winters, contribuíram para o desenvolvimento deste método, em que diferia da ideia de Brown, no alisamento da tendência e da sazonalidade. Assim, o seu método é designado método de Holt Winters.

Os métodos de alisamento exponencial são fundados em combinações ponderadas de observações do passado, em que as observações mais recentes recebem relativamente mais peso do que as observações mais antigas. Assim, os pesos atribuídos decrescem à medida que as observações são mais antigas (Hyndman et al., 2008).

Caso os dados não apresentem sazonalidade, estes apresentarão mais correlação com os dados consecutivamente anteriores, do que com os dados mais distantes. Por outro lado, no caso de dados sazonais, os mesmos correlacionam-se mais com as observações dos momentos homólogos mais próximos. (ver Caiado, 2016, pág.101).

Dentro dos métodos de alisamento exponencial, pode-se destacar o método de alisamento exponencial simples, o método de Holt e o método de Holt Winters. O último mencionado irá ser abordado no ponto 3.2.2.1, sendo um dos métodos utilizados no caso de estudo.

### 3.2.2.1 Método de Holt-Winters

O método de Holt-Winters, concebido por Holt e Winters, é um método de alisamento exponencial direcionado para prever dados com sazonalidade e tendência.

Compreende 3 componentes – o nível ( $L_t$ ), a tendência ( $T_t$ ) e a sazonalidade ( $S_t$ ) – e 3 parâmetros de alisamento – o  $\alpha$ ,  $\beta$  e o  $\gamma$ . O  $\alpha$  representa a constante de amortecimento do nível, o  $\beta$  é a constante de amortecimento da tendência e o  $\gamma$  é a constante de amortecimento do fator sazonal.

Quando o  $\alpha$  é pequeno, o peso dado às observações mais antigas decrescem mais lentamente. Quando o  $\alpha$  é grande, as observações mais antigas têm um peso muito pequeno, e, por isso, as resultantes séries são mais reativas às alterações recentes.

Pode-se verificar ainda que quando  $\alpha = 0$ , o nível é constante ao longo do tempo; quando  $\beta = 0$ , a tendência é constante ao longo do tempo; e quando  $\gamma = 0$ , o padrão sazonal é constante ao longo do tempo. (Hyndman et al., 2008).

As equações de nível, da tendência e da sazonalidade estão expressas em (5), (6) e (7), respetivamente. O  $m$  é a frequência da componente sazonal.



O método de Holt Winters compreende dois tipos: o método de Holt Winters aditivo e o método de Holt Winters multiplicativo. O método aditivo, expresso em (4) é preferencialmente utilizado para prever séries com sazonalidade constante ao longo do tempo. A sazonalidade exprime-se em termos absolutos e em cada ano varia em torno de 0. O ajustamento de série é realizado através da subtração de sazonalidade. (Hyndman & Athanasopoulos, 2018).

$$F_{t+h|t} = L_t + hT_t + S_{t+h-m(k+1)} \quad (4)$$

$$L_t = \alpha(Y_t - S_{t-m}) + (1 - \alpha)(L_{t-1} + T_{t-1}) \quad (5)$$

$$T_t = \beta * (L_t - L_{t-1}) + (1 - \beta *)T_{t-1} \quad (6)$$

$$S_t = \gamma(Y_t - L_{t-1} - T_{t-1}) + (1 - \gamma)S_{t-m} \quad (7)$$

O  $F$  representa a previsão da série com  $h$  períodos seguintes, o  $k$  representa a parte integrante de  $\frac{(h-1)}{m}$ , o que garante que as estimativas dos índices sazonais usados para previsão sejam provenientes do último ano da amostra (Hyndman & Athanasopoulos, 2018).

De acordo com os mesmos autores, a sazonalidade dos dados altera-se de forma proporcional é mais adequado a utilização do modelo multiplicativo (8) (Hyndman & Athanasopoulos, 2018). A sazonalidade (11) exprime-se em valores em torno de 1 e a sua soma é próxima de  $m$ , em cada ano (Hyndman & Athanasopoulos, 2018). Em vez da subtração, o método multiplicativo divide a série pela componente sazonal, para o seu ajustamento. (Hyndman & Athanasopoulos, 2018). As restantes componentes – nível e tendência – são expressas em (9) e (10), respetivamente.

$$F_{t+h|t} = (L_t + hT_t)S_{t+h-m(k+1)} \quad (8)$$

$$L_t = \alpha \left( \frac{Y_t}{S_{t-m}} \right) + (1 - \alpha)(L_{t-1} + T_{t-1})$$

(9)

$$T_t = \beta * (L_t - L_{t-1}) + (1 - \beta *)T_{t-1}$$

(10)

$$S_t = \gamma \left( \frac{Y_t}{(L_{t-1} + T_{t-1})} \right) + (1 - \gamma)S_{t-m}$$

(11)

### 3.2.3 Método Autorregressivo Integrado de Médias Móveis – ARIMA

Os métodos de ARIMA, utilizados para séries temporais, são caracterizados pela sua precisão e solidez matemática (Contreras, Espínola, Nogales, & Conejo, 2003).

O modelo ARIMA, que se centra nas autocorrelações de dados, é um modelo autorregressivo integrado de médias móveis, e assim, é constituído pela componente autoregressiva (AR), pela componente de diferenciação (I) e pela componente de Médias Móveis (MA). Estes constituintes serão analisados seguidamente individualmente.

#### 3.2.3.1 Componente Autoregressiva (AR)

O termo “autoregressão” indica uma regressão da própria variável. De ordem  $p$ , a equação adequada é sugerida em (12), onde o valor observado em  $t$  é explicado por  $p$  valores observados previamente.

$$Y_t = \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \dots + \varphi_p Y_{t-p} + \varepsilon_t$$

(12)

O  $\varepsilon_t$  corresponde ao processo de ruído branco, que é estacionário, uma vez que são observações aleatórias independentes do tempo. Assim, e como indica

Cryer e Chan (2008), o valor atual de  $Y_t$  é uma combinação linear das  $p$  observações passadas mais recentes, com o ruído branco ( $\varepsilon_t$ ).

Assim, o ruído branco são valores aleatórios de uma distribuição fixa, geralmente normal, com média 0 e variância  $\sigma_\varepsilon^2$  (ver Box, Jenkins, Reinsel, & Ljung, 2016, pág. 7).

Há dois tipos de modelos: os modelos determinísticos e os modelos probabilísticos.

De acordo com Box et al. (2016), os modelos determinísticos são modelos que determinam eventos futuros, através do cálculo exato. No entanto, os mesmos autores afirmam que, devido aos eventos desconhecidos que possam ocorrer, é inviabilizado este cálculo exato.

Por outro lado, o modelo estocástico é um modelo probabilístico que calcula um valor próximo, entre dois limites específicos. Assim, uma série temporal  $y_1, y_2, \dots, y_n$  de  $n$  observações sucessivas é considerada uma realização amostral duma população infinita dessas séries temporais que poderia ter sido gerada pelo processo estocástico. (Box et al., 2016).

Existem dois tipos de propriedades dos modelos estocásticos para descrever uma série temporal: a estacionaridade e a não estacionaridade.

De acordo com Hiper e McLeod (1994), geralmente os modelos estocásticos estacionários são projetados de modo a que a média e a variância sejam independentes do tempo. Assim, tanto a média como a variância são constantes ao longo do tempo.

Assim, a estacionaridade não depende da função do tempo. Por outro lado, uma série temporal com sazonalidade ou tendência é não estacionária, uma vez que essas componentes afetam os valores observáveis ao longo do tempo (ver Hyndman & Athanasopoulos, 2018). Assim, o primeiro caso de estudo deste trabalho é uma série temporal não estacionária, uma vez que é sazonal e tem tendência.

Como apontam Box et al. (2016), alguns métodos que utilizam médias móveis exponencialmente ponderadas podem ser apropriados para determinadas séries não estacionárias.

A estacionaridade pode ser considerada forte/estrita ou fraca/de segunda ordem. De acordo com Hipel e McLeod (1994), a estacionaridade forte acontece quando a distribuição conjunta de qualquer conjunto possível de variáveis aleatórias do processo não é afetada pelo tempo.

Assim, a estacionaridade forte significa intuitivamente que os gráficos em dois intervalos iguais de tempo de uma realização da série temporal devem exibir características estatísticas semelhantes. (ver Brockwell & Davis, 1990, pág. 12).

A estacionaridade fraca terá uma média e uma variância constantes e a covariância e a correlação serão funções da diferença de tempo. (ver Hillmer & Wei, 1991, pág.8)

### 3.2.3.2 Componente de Médias Móveis (MA)

Um processo de média móvel (MA) de ordem  $q$  é uma combinação linear do ruído branco atual com os termos  $q$  mais recentes do ruído branco histórico e é definido por (13) (Cowpertwait & Metcalfe, 2009).

$$Y_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \dots - \theta_q \varepsilon_{t-q} \quad (13)$$

De acordo com Cryer e Chan (2008), o  $Y_t$  é obtido ao aplicar os pesos  $1, -\theta_1, -\theta_2, \dots, -\theta_q$  às variáveis  $\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, \dots, \varepsilon_{t-q}$  e, de seguida, movendo os pesos e empregando-os a  $\varepsilon_{t+1}, \varepsilon_t, \varepsilon_{t-1}, \dots, \varepsilon_{t-q+1}$  para obter  $Y_{t+1}$  e assim por diante. Assim, a observação atual corresponde à aplicação dos pesos no ruído branco da série temporal para prever os próximos valores.

### 3.2.3.3 Método Autorregressivo de Médias Móveis – ARMA

O método autoregressivo de médias móveis é uma combinação entre o processo autoregressivo (AR) e o processo de médias móveis (MA), dando origem à sigla ARMA. A equação (14) indica o método ARMA  $(p, q)$ .

$$Y_t = \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \varphi_p Y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \quad (14)$$

### 3.2.3.4 Método Autorregressivo Integrado de Médias Móveis – ARIMA

O método Autorregressivo Integrado de Médias Móveis - ARIMA  $(p, d, q)$ , como já referido anteriormente tem 3 componentes: a componente autoregressiva (AR), a componente de diferenciação (I) e a componente de médias móveis (MA). Assim, o  $p$  é a ordem da parte autoregressiva,  $d$  é o grau de diferenciação e  $q$  é a ordem da parte da média móvel.

Este método, como indicam Brockwell e Davis (2016), é uma generalização do processo ARMA que inclui séries não estacionárias. Surge incluído aqui o termo da componente de integração, isto é, da diferenciação.

Considera-se que a diferenciação poderá ser representada em (15).

$$\Delta Y_t = Y_t - Y_{t-1} \quad (15)$$

A diferenciação, então, é a diferença entre a observação atual e a observação anterior. Caso a diferenciação de primeira-ordem não seja suficiente para estabilizar a média da série temporal, poder-se-á diferenciar a série  $d$  vezes, até a média estar estabilizada (Brockwell e Davis, 2016).

Por exemplo, caso a diferenciação seja de primeira-ordem, isto é, caso se pretenda aplicar o método ARIMA  $(p, 1, q)$ , é descrito em (16).

$$Y_t - Y_{t-1} = \varphi_1(Y_{t-1} - Y_{t-2}) + \varphi_2(Y_{t-2} - Y_{t-3}) + \cdots + \varphi_p(Y_{t-p} - Y_{t-p-1}) + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \quad (16)$$

Ainda, de forma a que o resultado seja uma observação em  $t$ , pode-se adequar (17).

$$Y_t = (1 + \varphi_1)Y_{t-1} + (\varphi_2 - \varphi_1)Y_{t-2} + (\varphi_3 - \varphi_2)Y_{t-3} + \cdots + (\varphi_p - \varphi_{p-1})Y_{t-p} - \varphi_p Y_{t-p-1} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \theta_q \varepsilon_{t-q} \quad (17)$$

### 3.2.3.5 Método Autorregressivo de Médias Móveis Sazonal - SARIMA

Para além dos modelos já apresentados, existe o SARIMA - Método Autorregressivo de Médias Móveis Sazonal. O método SARIMA adequa-se ao método aplicado no caso de estudo presente neste trabalho.

Assim como no modelo não sazonal, o modelo sazonal de ARIMA tem as componentes  $(p, d, q)$ , mas acrescenta ainda  $(P, D, Q)m$ , onde  $p$  é a ordem da parte autoregressiva,  $d$  é o grau de diferenciação e  $q$  é a ordem da parte da média móvel, e o  $m$  o número de observações por ano. As letras minúsculas indicam a não sazonalidade e as letras maiúsculas indicam a sazonalidade do método. Assim, o método SARIMA é representado como SARIMA  $(p, d, q)(P, D, Q)m$ .

A componente autoregressiva AR( $P$ ) é descrita em (18).

$$Y_t = \varphi_1 Y_{t-m} + \varphi_2 Y_{t-2m} + \cdots + \varphi_P Y_{t-Pm} + \varepsilon_t \quad (18)$$

A componente autoregressiva MA( $Q$ ) é descrita em (19).

$$Y_t = \varepsilon_t - \theta_1 \varepsilon_{t-m} - \theta_2 \varepsilon_{t-2m} - \cdots - \theta_Q \varepsilon_{t-Qm} \quad (19)$$

No fundo, como se pode verificar, os processos sazonais são idênticos aos descritos anteriormente para o método não sazonal, com a incorporação do  $m$  e das componentes sazonais, designadas como  $P$ ,  $D$  e  $Q$ .

### 3.2.3.6 Função de Autocorrelação

A autocorrelação pode ser descrita como uma correlação de dados consecutivos numa série temporal, e é de curto prazo quando, num período

curto, se relacionam dados num modelo com sazonalidade e tendência. (Chatfield, 2000)

De acordo com Hyndman e Athanasopoulos (2018), a função de autocorrelação mede a relação entre  $Y_t$  e  $Y_{t-k}$ , por diferentes valores de  $k$ . Assim, a função de autocorrelação indica se há um fraco ou forte relacionamento entre os dados históricos da série e o seguinte valor a prever. (ver Hyndman, Wheelwright, & Makridakis, 1998, pág. 41). Estes autores apontam para a sua adequação em indicar movimentos padronizados nos erros ou resíduos, em conjuntos de dados com um modelo de previsão aplicado, e em indicar se o método necessita de melhorias. A função é descrita em (20), onde  $\bar{Y}$  é a média das observações e o  $k$  os desfasamentos.

$$\frac{\sum_{t=k+1}^n (Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^n (Y_t - \bar{Y})^2}$$

(20)

### 3.2.4 Modelos de Regressão

Os modelos de regressão são modelos matemáticos que descrevem uma relação entre uma variável dependente e uma ou mais variáveis independentes.

Há vários tipos de modelos de regressão: a regressão linear simples, a regressão linear múltipla, regressão não linear, regressão polinomial, regressão logística, regressão de ridge, regressão de lasso, regressão elasticnet, entre outros.

No segundo caso de estudo, foram implementados vários modelos de regressão, os quais vão ser descritos ao longo desta secção.

### 3.2.4.1 Modelo de Regressão Linear Simples

O modelo de regressão linear simples modelo uma relação entre 2 variáveis: uma independente (X) e uma dependente (Y). O modelo é descrito em (21).

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

21

O  $\beta_0$  e  $\beta_1$  são dois parâmetros desconhecidos que representam os termos de interceção e inclinação no modelo linear. (James, Witten, Hastie, & Tibshirani, 2013).

De acordo com James et al. (2013), um dos métodos mais comumente utilizados para estimar os parâmetros é o método dos mínimos quadrados, em que se minimiza a soma dos resíduos quadrados.

A soma dos resíduos quadrados (SRQ) pode ser descrita em (22).

$$SRQ = (Y_1 - \hat{\beta}_0 - \hat{\beta}_1 X_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 X_2)^2 + \dots + (y_t - \hat{\beta}_0 - \hat{\beta}_1 X_t)^2$$

22

Onde o acento circunflexo representa a estimativa dos parâmetros  $\beta_0$  e  $\beta_1$ .

A minimização da soma dos resíduos quadrados permite determinar o  $\hat{\beta}_0$  e o  $\hat{\beta}_1$ , dadas pelas equações (23) e (24), respetivamente.

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

23

$$\hat{\beta}_1 = \frac{\sum_{t=1}^T (X_t - \bar{X})(Y_t - \bar{Y})}{\sum_{t=1}^T (X_t - \bar{X})^2}$$

24

De outra forma e equivalente a (24), o  $\hat{\beta}_1$  pode ser descrito em (25).

$$\hat{\beta}_1 = \frac{Cov(X, Y)}{Var(X)}$$

25

Assim, os parâmetros do modelo definem a reta dos quadrados mínimos (James et al., 2013).



De modo a corroborar a proximidade da realidade da estimativa dos parâmetros do modelo é proposto a verificação dos erros padrão destes parâmetros (James et al., 2013). Assim, pode-se utilizar as equações descritas em (26) e (27), que representam o erro-padrão (EP) da estimativa de  $\beta_0$  e o erro-padrão da estimativa de  $\beta_1$ , onde  $\sigma^2$  representa a variância de  $\varepsilon$ . Segundo James et al. (2013), o  $\sigma^2$  dos erros ( $\varepsilon$ ) para cada observação não são correlacionados com a variância comum e, o mesmo, pode ser estimado através do erro-padrão residual (28).

$$EP(\hat{\beta}_0)^2 = \sigma^2 \left[ \frac{1}{T} + \frac{\bar{X}^2}{\sum_{t=1}^T (X_t - \bar{X})^2} \right]$$

26

$$EP(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{t=1}^T (X_t - \bar{X})^2}$$

27

$$EPR = \sqrt{SRQ/(t - 2)}$$

28

Como James et al. (2013) indicam, os erros-padrão podem ser utilizados para construir intervalos de confiança e testes de hipóteses.

Os intervalos de confiança para os parâmetros do modelo podem ser identificados em (29) e (30). Neste ponto, os autores James et al. (2013) referenciam que são fórmulas aproximadas, uma vez que é pressuposto que os erros sejam gaussianos, e, nas fórmulas dos erros-padrão dos parâmetros, o número 2 irá variar, dependendo do número de observações do modelo linear e com o nível de significância.

$$IC\hat{\beta}_0 = \hat{\beta}_0 \pm 2 \cdot EP(\hat{\beta}_0)$$

29

$$IC\hat{\beta}_1 = \hat{\beta}_1 \pm 2 \cdot EP(\hat{\beta}_1)$$

30

Geralmente, no âmbito dos testes de hipóteses, há uma hipótese nula e uma hipótese alternativa. A hipótese nula sugere que não há relação entre as

variáveis  $Y$  e  $X$ , isto é,  $H_0 : \beta_1 = 0$ . A hipótese alternativa, pelo contrário, alvitra uma relação entre as variáveis  $Y$  e  $X$ , correspondendo a  $H_1 : \beta_1 \neq 0$  (ver James et al., 2013, pág. 67).

A medida para se obter o número de desvios-padrão em que  $\hat{\beta}_1$  se afasta de 0 é denominada de *t-statistic* (31) (James et al., 2013). Deste modo, *t-statistic* permite a refutação ou a confirmação das hipóteses.

$$t - statistic = \frac{\hat{\beta}_1 - 0}{EP(\hat{\beta}_1)}$$

31

Um termo associado a esta problemática é o *p-value*. O *p-value*, representa a probabilidade de se obter o valor de *t-statistic* no caso da hipótese nula ser verdadeira (James et al., 2013). Caso o *p-value* seja alto, isto é, se for superior a 1% ou 5%, a hipótese nula não é rejeitada.

Os valores 1% e 5% são os valores comumente utilizados para a determinação da relação ou não-relação entre as variáveis (James et al., 2013).

Assim, a regressão linear poderá ser descrita em (32).

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

32

#### 3.2.4.2 Modelo de Regressão Linear Múltipla

O modelo de regressão linear múltipla descreve uma relação entre a variável dependente ( $Y$ ) e duas ou mais variáveis independentes ( $X$ ).

Assim, o conjunto das variáveis independentes podem ser descritas como  $X = X_1, X_2, \dots, X_n$  e  $\beta = \beta_0, \beta_1, \dots, \beta_n$  representa a lista de parâmetros correspondente a  $X_1, \dots, X_t$  (Harrell, 2015).

Pode-se descrever a regressão linear múltipla em (33).

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n + \varepsilon$$

33

A previsão de  $Y$  no modelo de regressão linear múltipla (34) é composta pela estimativa dos parâmetros do modelo, as variáveis independentes e o resíduo. Podemos verificar que é similar ao modelo de regressão linear simples, distinguindo-se pela adição de variáveis independentes.

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \dots + \hat{\beta}_n X_n$$

34

Tal como a regressão linear simples, pode ser aplicado a soma dos resíduos quadrados para a estimativa destes parâmetros. Neste contexto, a soma dos resíduos quadrados é expressa em (35), onde  $n$  representa as variáveis independentes e  $t$  a observação.

$$SRQ = \sum_{t=1}^T (Y_t - \hat{\beta}_0 - \hat{\beta}_1 X_{1t} - \hat{\beta}_2 X_{2t} - \dots - \hat{\beta}_n X_{nt})^2$$

35

Segundo James et al. (2013), como o modelo de regressão linear múltiplo tem diversas variáveis independentes ( $n$ ), na hipótese nula no teste de significância global, verifica-se se todos os coeficientes se igualam a 0. Assim, em vez de uma variável, serão várias variáveis tidas em conta, isto é,  $H_0: \beta_1 = \beta_2 = \dots = 0$ . A hipótese alternativa será que, pelo menos uma variável não se igual a 0, ou seja,  $H_1: \text{pelo menos uma } \beta_n \neq 0$ , onde  $\beta_n$  é interpretado como uma média do efeito em  $Y$  do aumento de uma unidade de  $X_n$ , mantendo-se as outras variáveis fixas (ver James et al., 2013, pág.72).

Quando se aborda este modelo, é importante referir a seleção de variáveis oportunas para o modelo, isto é, quais variáveis a incluir no modelo e como o fazer (James et al., 2013). Os mesmos autores propõem e explicam 3 formas clássicas de o realizar: a seleção *upward*, a seleção *backward* e a seleção mista.

Como os mesmo autores elucidam, a seleção *upward* caracteriza-se por, primeiramente, adicionar ao modelo a variável com o valor mais baixo da soma dos resíduos quadrados (SRQ). Seguidamente, inclui a segunda variável

com menos SRQ, e assim em diante, até o acréscimo de variáveis não seja significativo e adequado para o modelo. Pelo contrário, a seleção *backward*, com todas as variáveis no modelo, retira a variável com o maior *p – value*. Assim, o processo continua até não ser estatisticamente benéfico a remoção das variáveis (James et al., 2013).

Como explicam James et al. (2013), o processo de seleção mista inicia-se com um modelo sem variáveis em que se vai acrescentando variáveis, uma a uma. Quando uma variável excede um determinado limite do *p – value*, retiramos a mesma. Assim, vai-se acrescentando e retirando variáveis, de acordo com o melhor para o modelo (James et al., 2013).

#### 3.2.4.3 Modelo de Regressão de Lasso

O modelo de regressão de Lasso - Least Absolute Shrinkage and Selection Operator – é um método de encolhimento e de seleção de variáveis.

O modelo utiliza uma técnica que restringe ou regulariza as estimativas do coeficiente, ou equivalentemente, que reduz as estimativas do coeficiente para zero, o que pode reduzir significativamente a sua variância (James, Witten, Hastie, & Tibshirani, 2013, pág. 215).

A sua popularidade deve-se, em parte, a uma característica principal do procedimento: encolhimento do vetor de coeficientes de regressão para 0, com a possibilidade de definir alguns coeficientes idênticos a 0, resultando num procedimento simultâneo de estimativa e seleção de variáveis. (Hans, 2009)

Deste modo, este modelo tem a particularidade de excluir variáveis não significativas para o modelo, ao contrário do Modelo de Regressão de Ridge. Assim, os coeficientes podem-se igualar a 0, desde que o seu parâmetro de ajustamento ( $\lambda$ ) seja suficientemente grande (James, Witten, Hastie, & Tibshirani, 2013).

O modelo de regressão de Lasso (38) é constituído por 2 componentes: pela soma dos resíduos quadrados (SRQ), descrita em (37), e pela designada penalidade de encolhimento. A penalidade de encolhimento inclui o parâmetro de ajustamento ( $\lambda$ ), que pode ir desde 0 a  $\infty$ , sendo definido por *cross validation* (James, Witten, Hastie, & Tibshirani, 2013).

A soma dos resíduos quadrados (SRQ), já mencionada nos modelos anteriores, é designada por (37).

$$SRQ = \sum_{t=1}^T (y_t - \beta_0 - \sum_{n=1}^N \beta_n x_{tn})^2$$

36

$$Modelo de Regressão de Lasso = SRQ + \lambda \sum_{n=1}^N |\beta_n|$$

37

#### 3.2.4.4 Modelo de Regressão de Ridge

O modelo de Regressão de Ridge (40) é composto pela soma dos resíduos quadrados (SRQ), descrita em (39), e por uma penalidade de encolhimento, fazendo parte o parâmetro de ajustamento ( $\lambda$ ).

O parâmetro de ajustamento ( $\lambda$ ) é superior ou igual a 0.

$$SRQ = \sum_{t=1}^T (Y_t - \beta_0 - \sum_{n=1}^N \beta_n x_{tn})^2$$

38

$$Modelo de Regressão de Ridge = SRQ + \lambda \sum_{n=1}^N \beta_n^2$$

39

Ao contrário do modelo de regressão de Lasso, a penalidade de encolhimento irá encolher os coeficientes para próximos de 0, mas nunca igual a 0, a não ser que  $\lambda = \infty$  (James et. Al, 2013). Assim, não exclui nenhuma variável do seu modelo.

### 3.2.4.5 Modelo de Regressão da Floresta Aleatória

De acordo com James et al. (2013), a agregação de bootstrap, ou bagging, é um procedimento de uso generalizado para reduzir a variância de bagging de um método de aprendizagem estatístico.

Como explicam James et al. (2013), o modelo de regressão da floresta aleatória, é contruído um conjunto de árvores de decisão em amostras de treino de bootstrap. Neste modelo, efetua-se a divisão de árvores, e cada vez que isto se verifica, é considerado uma amostra aleatória de  $m$  preditores do conjunto completo de variáveis independentes ( $n$ ). Assim, a cada divisão de uma árvore, o modelo não considera todas as variáveis independentes, o que resulta numa menor variância (James et al., 2013).

Como James et al. (2013) explana, a cada divisão é feita uma nova amostra de  $m$  preditores, em que, geralmente,  $m$  é aproximadamente a raiz quadrada de  $n$ .

## 3.3 Métricas de Precisão

O erro de previsão é a diferença entre o valor observado da série e o valor estimado da série, num determinado instante, como se pode verificar em (41).

$$e^t = Y^t - F^t \quad (40)$$

Onde  $e_t$  é o erro de previsão no tempo  $t$ ,  $Y_t$  representa o valor observado no tempo  $t$  e o  $F_t$  é o valor de previsão no tempo  $t$ .

Há várias medidas para agregar os erros em (41) em medidas sumárias da qualidade e/ou enviesamento de previsão. Por exemplo, o erro médio (ME - Mean Error) é a média dos erros de cada período  $t$  analisado, onde  $T$  representa o número de observações disponíveis para o teste (e.g. Hyndman, Wheelwright e Makridakis, 1998).

$$ME = \frac{1}{T} \times \sum_{t=1}^T e_t \quad (41)$$

Esta medida de erro tem o problema de erros positivos e negativos se anularem, o que significa que um valor reduzido do ME, não significa que o modelo seja bom. Para evitar este problema, os erros devem ser considerados sem sinal. As duas formas mais comuns de agregar erros é somando o seu módulo ou o seu quadrado (em ambos os casos o sinal é ignorado). As medidas de erro mais comuns são o erro quadrático médio (MSE - Mean Square Error), e a raiz quadrada do erro quadrático médio (RMSE – Root Mean Square Error) e o erro absoluto médio (MAE – Mean Absolute Error) (ver Hyndman e Koehler, 2006).

$$\text{MSE} = \frac{1}{T} \times \sum_{t=1}^T e_t^2$$

(42)

$$\text{RMSE} = \sqrt{\text{MSE}}$$

(43)

$$\text{MAE} = \frac{1}{T} \times \sum_{t=1}^T |e_t|$$

(44)

$$\text{MAPE} = \frac{1}{T} \times \sum_{t=1}^T \left| \frac{e_t}{Y_t} \right| * 100$$

(45)

O MSE apresenta-se em unidades de medida ao quadrado. O MSE é empregue regularmente, devido à sua facilidade de manuseamento matemático. Por outro lado, o RMSE, como é a raiz quadrada do MSE, encontra-se na mesma medida que os dados e, por isso, mais favorável de ser utilizada. (ver Hyndman & Koehler, 2006, pág. 682).

De acordo com Hyndman e Koehler (2006), tanto a RMSE como a MSE desempenham um papel importante no âmbito dos modelos estatísticos, devido maioritariamente à sua fundamentação teórica.

Outra medida utilizada é a MAE, representada em (45), que transforma os erros em erros positivos para a realização da sua média. É, assim, uma fórmula de mais fácil interpretação e mais fácil explanação a não-especialistas (Hyndman, Wheelwright, & Makridakis, 1998, pág. 43).

De acordo com Hyndman e Koehler (2006), nas medidas RMSE e MSE há mais suscetibilidade aos outliers comparativamente à métrica MAE.

A MAPE é apresentada em percentagem, e por isso, é vantajosa para comparar diferentes métodos de previsão. No entanto, de acordo com Hyndman e Koehler (2006), quando avalia dados iguais ou próximos de 0, os resultados são inadequados, pois não são finitos ou não são definidos, ou são enviesados. A relevância que é dada ao 0 por parte da MAPE também resulta numa desvantagem (por exemplo “não fazem sentido ao medir o erro de previsão para temperaturas nas escalas Fahrenheit ou Celsius”) (Hyndman & Koehler, 2006, pág. 683).

Hyndman e Koehler elaboraram uma nova medida, o erro médio absoluto em escala (MASE – Mean Absolute Scaled Error). A MASE, descrita em (48), é baseada nos erros em escala (47). É a média absoluta da divisão do erro no instante  $t$ , pelo erro absoluto médio do método naïve. Ora veja-se que, o erro cometido pelo método naïve é a diferença entre o valor observado no tempo  $i$  e o valor observado anteriormente no tempo  $i$  (ver (49)), uma vez que no método naïve o valor observado no momento de tempo  $t - 1$  é a previsão para o momento  $t$ . Note-se que o  $i$  também é um índice de tempo, tal como  $t$ .

Os criadores da MASE defendem que a mesma é melhor comparativamente às restantes, já que lida bastante bem com escalas díspares, sem ter problemas com dados próximos de 0 ou negativos, como a MAPE tem, e mostra sempre um resultado definido e finito, exceto quando todos os dados são uniformes. Ainda, os mesmos autores afirmam que a sua interpretação é fácil, pois caso a medida seja maior que 1, quer dizer que é pior que o método naïve. Por



exemplo, caso o erro absoluto em escala seja de 20%, quer dizer que o método utilizado é 80% melhor que o método naïve.

$$q_t = \frac{e_t}{\frac{1}{t-1} \sum_{i=2}^T |Y_i - Y_{i-1}|}$$

(46)

$$\text{MASE} = \text{mean}(|q_t|)$$

(47)

$$\text{Erro do Método Naïve} = Y_i - Y_{i-1}$$

(48)

De notar que, embora a MASE resolva alguns problemas, a MAE poderá ser preferível quando os dados se encontram na mesma escala, e a MAPE quando os dados não são negativos e são afastados do 0. (Hyndman & Koehler, 2006,pág. 687).

No âmbito dos modelos de regressão, podemos designar várias medidas de precisão, para a verificação da adequação dos modelos aos dados em questão.

O erro-padrão dos resíduos (50) do modelo de regressão linear simples, abordado em 4.5.1, assinala a estimativa do desvio-padrão dos erros ( $\varepsilon$ ).

$$EPR = \sqrt{SRQ/(t-2)}$$

49

Por outro lado, o erro-padrão dos resíduos para o modelo de regressão linear múltiplo é descrito em (51), onde  $n$  representa o número de variáveis.

$$EPR = \sqrt{\frac{1}{t-n-1} SRQ}$$

50

Uma medida comumente utilizada em estatística é o  $R^2$ (52), o qual expressa o quanto Y é explicado pela variável X.

$$R^2 = \frac{STQ - SRQ}{STQ} = 1 - \frac{SRQ}{STQ}$$

51

Onde STQ é a soma total dos quadrados e é descrita em (53).

$$STQ = \sum (y_t - \bar{y})^2$$

52

## Capítulo 4

### 4. Aplicação do Caso de Estudo

#### 4.1 Vendas da Empresa “OLI – Sistemas Sanitários, S.A.”

A fim de analisar e perspetivar os modelos preditivos mais adequados e eficazes para a previsão dos dados, foi utilizado o programa R.

Como Bernard afirma, “não existe uma boa gestão sem boa previsão de vendas”, por isso, a aplicabilidade preditiva de dados destinou-se à obtenção de informações das próximas vendas totais mensais da empresa “OLI – Sistemas Sanitários, S.A”. Para este alcance, foram recolhidos dados desde o mês de janeiro de 2002 até ao mês de outubro de 2019, com o objetivo de obter valores estimados até fevereiro de 2020.

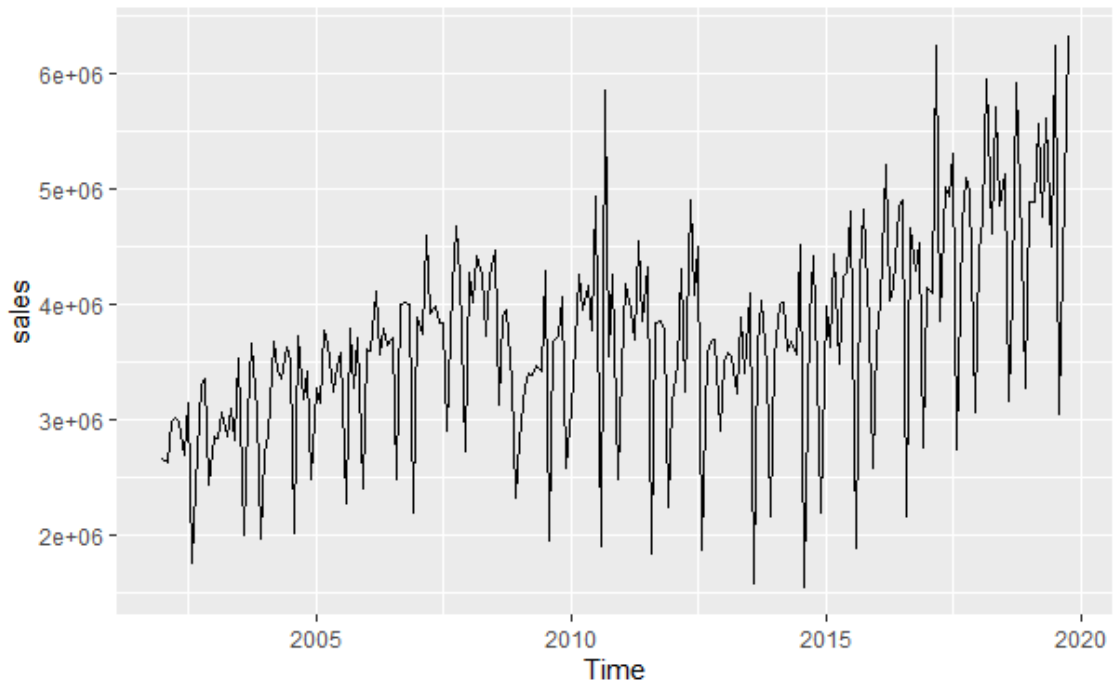
Em primeiro lugar, usou-se livreria xlsx para extrair os dados do Excel para o R:

```
my_table <- read_excel("Dados.xlsx")  
View(my_table)
```

Depois da introdução do conjunto de dados das vendas da empresa no R, foi analisada graficamente o mesmo, com a função autoplot. Este gráfico (figura 2) representa as vendas de cada mês no eixo do y, e o eixo do x o período temporal.

Figura 2

Gráfico Temporal das Vendas Totais



O gráfico da figura 2, foi obtido através do código:

```
sales <- ts(my_table[,2], start = c(2002,1), end = c(2019,10), frequency = 12)
autoplot(sales)
```

Note-se que a primeira linha do código com a função `ts`, permite a criação dos dados em séries temporais.

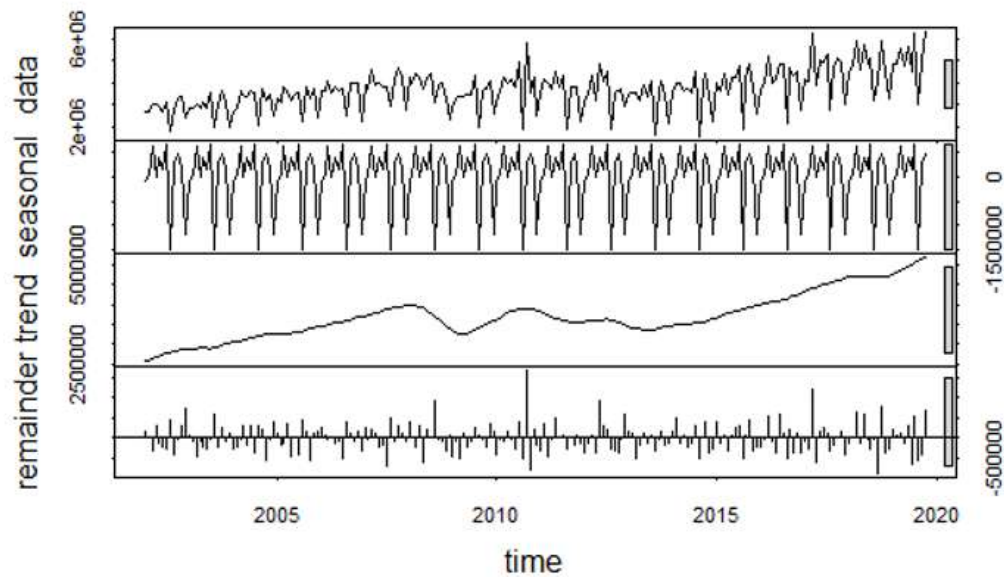
Com a função `STL` (Seasonal and Trend decomposition using Loess), um método concebido por Cleveland, McRae e Terpenning, foi realizada a decomposição das componentes dos dados:

```
fit <- stl(sales, s.window="period")
plot(fit)
```

Assim, a figura 3 demonstra a análise da sua sazonalidade, da sua linha tendência e dos seus *outliers*, visando uma melhor compreensão e interpretação da informação.

Figura 3

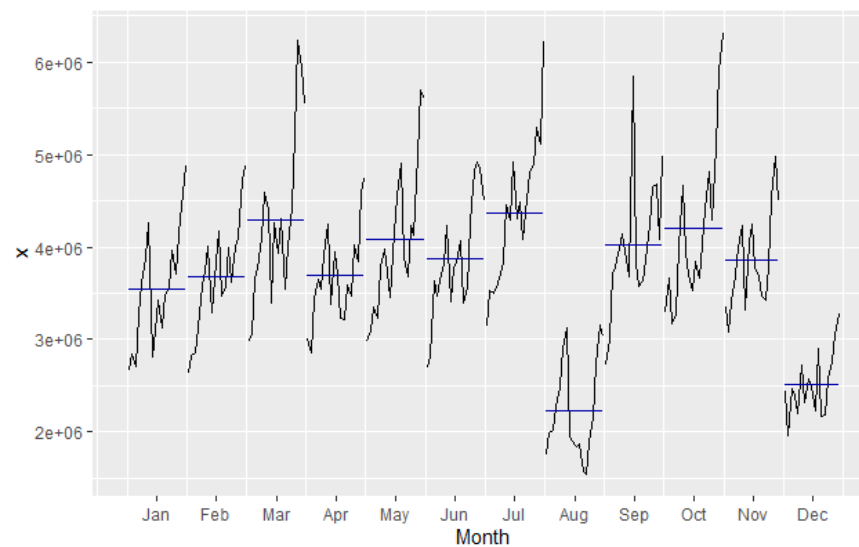
Gráfico de Dados, Sazonalidade, Tendência e Componentes Irregulares



Para uma observação de subséries temporais em cada estação, foi elaborado o gráfico da figura 4, em que a linha azul representa a média das observações de cada mês. Ou seja, a linha horizontal azul representa a média dos valores observados de cada mês de todos os anos considerados.

Figura 4

Subséries Temporais de cada Estação



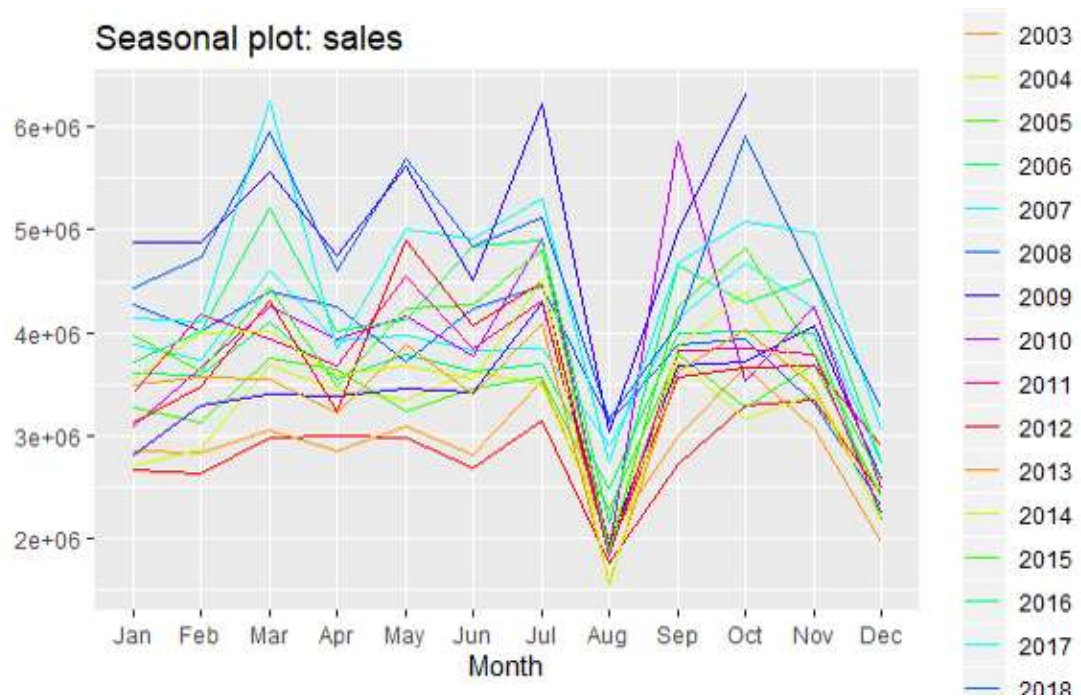
O gráfico da figura 4 foi obtido através do seguinte código:

```
ggmonthplot(sales)
```

Para uma exploração mais pormenorizada, foi elaborado um gráfico para analisar a sazonalidade – linhas diferenciadas para cada ano e os meses no eixo dos x (ver Figura 4). Na Figura 5 observa-se, num primeiro instante, e de forma holística, que as vendas têm vindo a aumentar, com exceção dos anos 2008 e 2009, e o período compreendido entre 2011 e 2014 (porque as linhas destes anos se encontram acima das restantes). É de destaque o acréscimo de montante nos meses de fevereiro, maio, julho, setembro e outubro, e um decréscimo acentuado no mês de agosto. Estes acréscimos e decréscimos são constantes no tempo, o que denota sazonalidade.

Figura 5

Sazonalidade por Mês e Ano



O gráfico da Figura 5 foi obtido através do código:

```
ggseasonplot(sales, col=rainbow(10),year.labels=FALSE, continuous = FALSE)
```

Posteriormente a esta fase preliminar de reconhecimento dos dados, foi dado o começo efetivo das previsões e das métricas de precisão, com os métodos preditivos Holt-Winters aditivo e multiplicativo e o método ARIMA.

A primeira metodologia utilizada foi o método de alisamento exponencial de Holt-Winters aditivo, com o seguinte código:

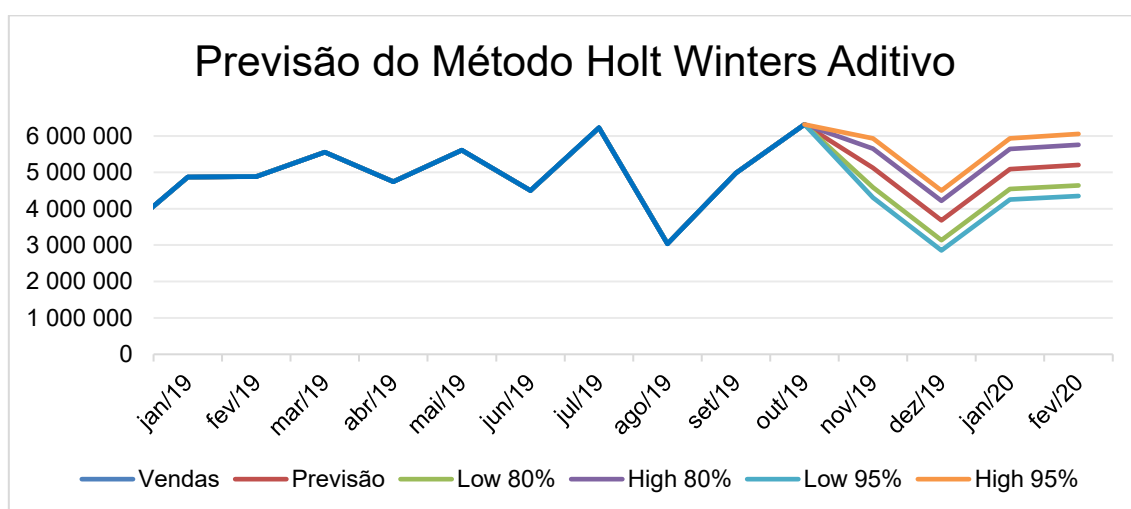
```
hw1 <- hw(sales, h = 4, seasonal = "additive")  
autoplot(hw1)
```

Onde “sales” representa os dados, “h” o número de períodos a ser estimado, que, neste caso, são quatro, isto é, de novembro de 2019 até fevereiro de 2020, e “seasonal” o método pretendido.

Através da função autoplot, foi obtido o gráfico da figura 6, que representa as vendas da “OLI – Sistemas Sanitários, S.A” e as vendas perspectivadas com o método Holt Winters aditivo, até fevereiro de 2020.

Figura 6

Gráfico da Previsão do Método Holt Winters Aditivo



De um modo geral, pode-se observar que o modelo prevê uma descida de vendas até dezembro de 2019, e uma constante subida até fevereiro de 2020 (ver tabela 1).

Tabela 1

Previsão das Vendas Totais da Empresa, com o Método Holt Winters Aditivo

Ano	Mês	Previsão	Low 80%	High 80%	Low 95%	High 95%
2019	Novembro	5 123 771	4 591 994	5 655 547	4310489	5 937 053
2019	Dezembro	3 677 890	3 137 793	4 217 987	2851883	4 503 898
2020	Janeiro	5 091 249	4 542 569	5 639 930	4252115	5 930 384
2020	Fevereiro	5 202 119	4 644 596	5 759 643	4 349 461	6 054 778



Os intervalos de confiança são calculados com 80% e 95% de confiança, para a observação mais minuciosa da previsão. Os intervalos de confiança são calculados através da fórmula descrita em (30).

$$F_{t+h|t} \pm c\hat{\sigma}_h$$

(53)

O  $\hat{\sigma}_h$  corresponde a uma estimação do desvio padrão em ordem à distribuição de previsão  $h$  e o  $c$  é o multiplicador. O valor do multiplicador dependerá do intervalo de confiança que se pretende estimar. Por exemplo, para o intervalo de confiança de 80% será 1,28 e para o de 95% será 1,96 (ver Hyndman & Athanasopoulos, 2018). Deste modo, as fórmulas para os intervalos descritos anteriormente são representadas em (55), para o intervalo de confiança de 80%, e (56), para o de 95%.

$$F_{t+h|t} \pm 1,28\hat{\sigma}_h$$

(54)

$$F_{t+h|t} \pm 1,96\hat{\sigma}_h$$

(55)

A função `summary` no R, para além de indicar as previsões, também fornece dados relativos aos parâmetros de alisamento. Assim, o  $\alpha$  é 0,1734, o  $\beta$  de 0,0042 e o  $\gamma$  de 0,2968. Isto significa que o  $\alpha$  é reduzido o que significa que o nível da série é pouco reativo e bastante alisado. O  $\beta$  e o  $\gamma$  também são valores pequenos, o que expressa um peso razoável dado aos valores históricos, enquanto que se fossem valores altos, revelava um peso pouco significativo ao passado.

No sentido de averiguar o nível de fiabilidade do método de Holt-Winters aditivo, foram analisadas as medidas de precisão definidas no capítulo 4.7, cujas são ME, RMSE, MAE, MAPE, MASE e ACF1 (função de autocorrelação de primeira-ordem).

Tabela 2

Métricas de Precisão da Previsão das Vendas Totais da Empresa, com o Método Holt Winters Aditivo

	ME	RMSE	MAE	MAPE	MASE	ACF1
<b>HW</b> <b>Aditivo</b>	-12 044.93	399 134	305 965.6	8.639406	0.7895211	-0.1870798

O método de Holt Winters multiplicativo foi obtido através das mesmas funções que o aditivo, com a diferença de se indicar multiplicativo, como se pode verificar:

```
hw2 <- hw(sales, h = 4, seasonal = "multiplicative")
autoplot(hw2)
```

Comparativamente ao método aditivo, o multiplicativo é mais pessimista, no sentido em apresenta valores mais baixos exceto no mês de novembro, como se pode ver na figura 7 e tabela 3. Encontra-se mais próximo dos dados reais no mês de novembro de 2019 e no mês de fevereiro 2020.

Figura 7

Gráfico da Previsão do Método Holt Winters Multiplicativo

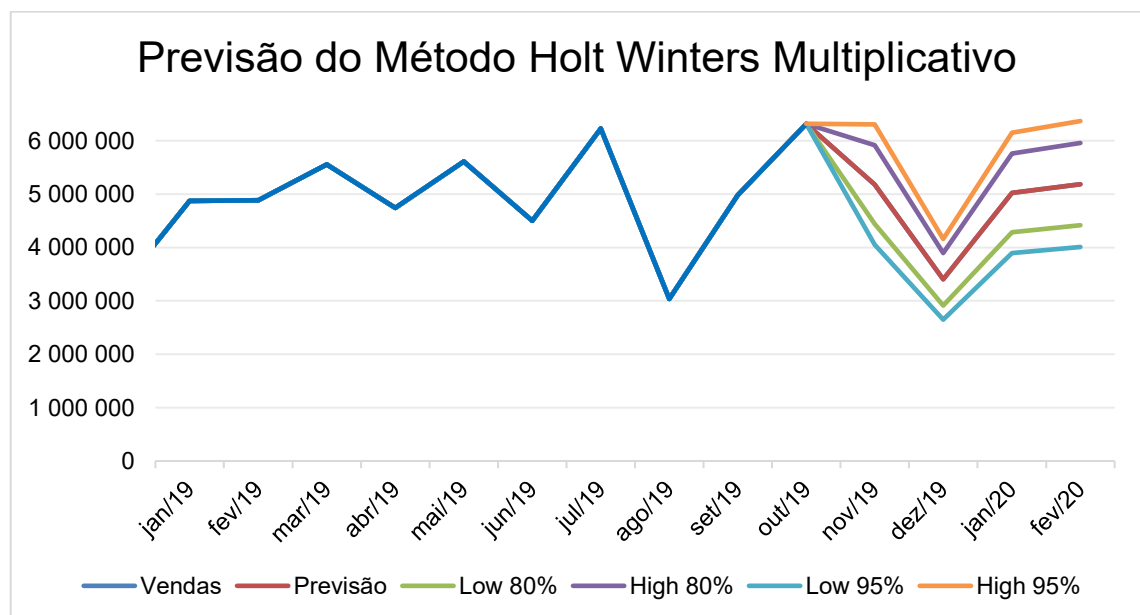


Tabela 3

Previsão das Vendas Totais da Empresa, com o Método Holt Winters Multiplicativo

<i>Ano</i>	<i>Mês</i>	<i>Previsão</i>	<i>Low 80%</i>	<i>High 80%</i>	<i>Low 95%</i>	<i>High 95%</i>
2019	Novembro	5 178 384	4 439 680	5 917 088	4 048 634	6 308 134
2019	Dezembro	3 403 411	2 910 838	3 895 985	2 650 085	4 156 738
2020	Janeiro	5 024 650	4 287 037	5 762 264	3 896 568	6 152 733
2020	Fevereiro	5 187 764	4 415 511	5 960 017	4 006 705	6 368 823

Todas as medidas de erro calculadas são inferiores em relação método aditivo (ver tabela 4).

Tabela 4

Métricas de Precisão da Previsão das Vendas Totais da Empresa, com o Método Holt-Winters Multiplicativo e o Método Holt-Winters Aditivo

	<i>ME</i>	<i>RMSE</i>	<i>MAE</i>	<i>MAPE</i>	<i>MASE</i>	<i>ACF1</i>
<i>HW</i> <i>Multiplic</i> <i>ativo</i>	-35 531,35	397 227.3	299 392.1	8.510605	0.7725588	-0.1850994
<i>HW</i> <i>Aditivo</i>	-12 044.93	399 134	305 965.6	8.639406	0.7895211	-0.1870798

Por outro lado, o  $\alpha$  é 0,1694, o  $\beta$  de 0,0014 e o  $\gamma$  de 0,2002.

No método ARIMA, verificamos que também apresenta a mesma oscilação do que nos outros métodos (ver figura 8).

Figura 8

Gráfico da Previsão do Método ARIMA

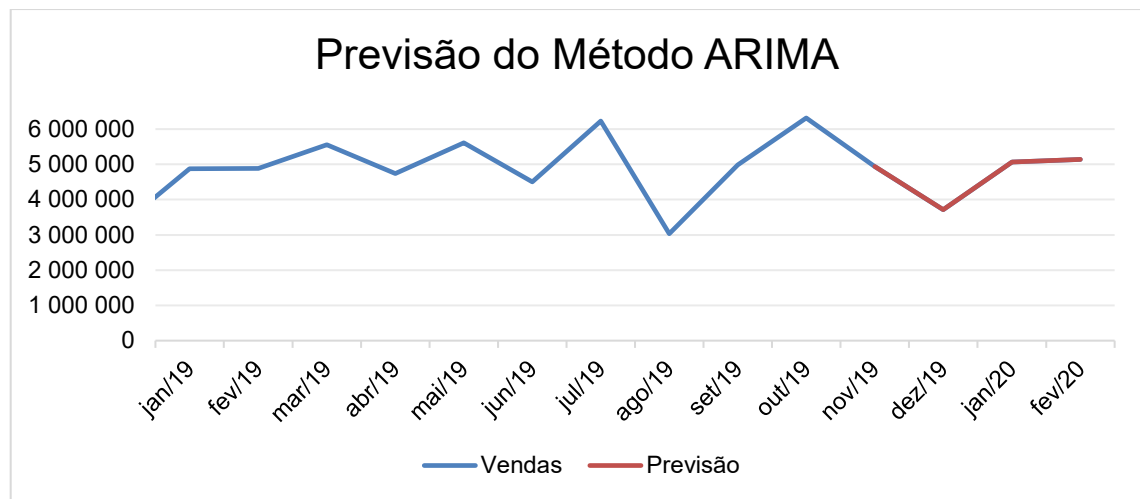


Tabela 5

Previsão das Vendas Totais da Empresa, com o Método ARIMA

<i>Ano</i>	<i>Mês</i>	<i>Previsão</i>
2019	Novembro	4 937 236
2019	Dezembro	3 716 084
2020	Janeiro	5 067 758
2020	Fevereiro	5 141 000

Tabela 6

Métricas de Precisão da Previsão das Vendas Totais da Empresa, com o Método ARIMA

	<i>ME</i>	<i>RMSE</i>	<i>MAE</i>	<i>MPE</i>	<i>MAPE</i>	<i>MASE</i>	<i>ACF1</i>
<b>ARIMA</b>	38 825.97	376 452.8	286 119.1	-0.1380788	7.914312	0.7383088	-0.003323674

Aquando esta previsão, os valores reais das vendas ainda não eram sabidos, uma vez que o objetivo era prever os mesmos. Depois de realizada esta análise, na tabela 7, verifica-se os números reais das vendas, no entanto, foram obtidos a posteriori.

Assim, e numa ótica de comparação (ver tabela 7), verificamos que as previsões nos métodos de Holt Winters estão na ordem dos 5 000 000, com

exceção do mês de dezembro de 2019. Já o modelo ARIMA, no mês de novembro não ultrapassa os 5 000 000, aproximando-se mais das vendas da OLI, como indica a tabela 7 e o gráfico da figura 9.

Figura 9

Gráfico de Comparação dos Métodos de Previsão

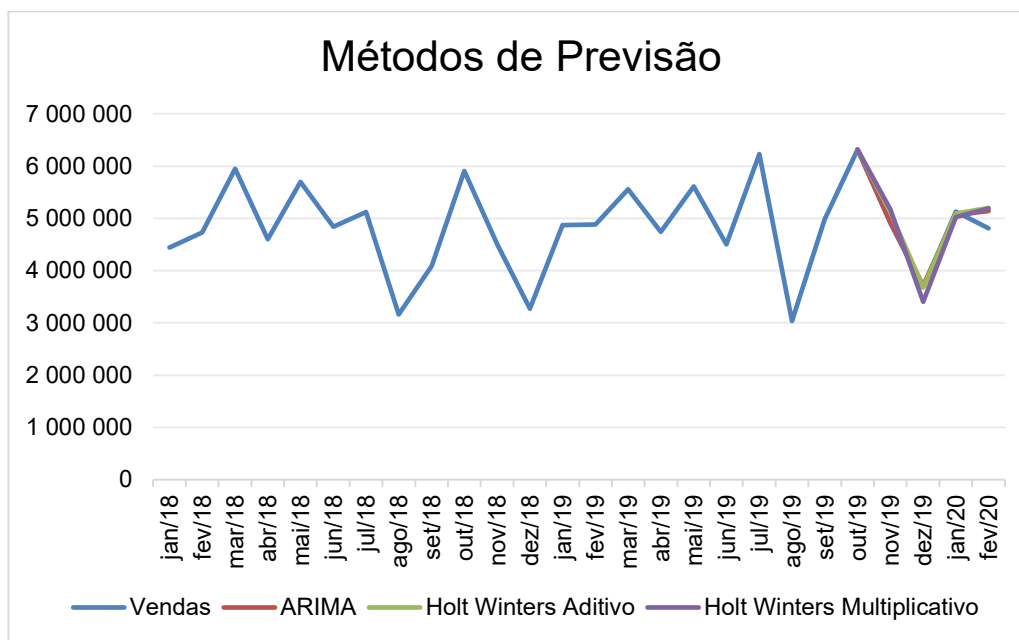


Tabela 7

Comparação dos Métodos de Previsão

<i>Ano</i>	<i>Mês</i>	<i>Real</i>	<i>Holt Winters Aditivo</i>	<i>Holt Winters Multiplicativo</i>	<i>ARIMA</i>
2019	Novembro	4 914 882	5 123 771	5 178 384	4 937 236
2019	Dezembro	3 684 854	3 677 890	3 403 411	3 716 084
2020	Janeiro	5 125 799	5 091 249	5 024 650	5 067 758
2020	Fevereiro	4 809 450	5 202 119	5 187 764	5 141 000

As medidas dos erros são inferiores no método ARIMA, comparativamente aos demais, salvo o ME, que se apresenta como positivo e maior (em termos absolutos) que nos restantes métodos (ver tabela 8).

Tabela 8

Comparação de Métricas de Precisão

	<i>ME</i>	<i>RMSE</i>	<i>MAE</i>	<i>MAPE</i>	<i>MASE</i>	<i>ACF1</i>
<i>HW Aditivo</i>	-12 044.93	399 134.0	305 965.6	8.639406	0.7895211	-0.1870798
<i>HW Multiplicativo</i>	-35 531,35	397 227.3	299 392.1	8.510605	0.7725588	-0.1850994
<i>ARIMA</i>	38 825.97	376 452.8	286 119.1	7.914312	0.7383088	-0.003323674

Após esta a análise preditiva e de precisão, podemos averiguar que o método que mais se adequa para as vendas da OLI – Sistemas Sanitários, S.A é o método de ARIMA, uma vez que foi o método que apresentou menores erros.

## 4.2 Consumos Energéticos da Empresa “OLI – Sistemas Sanitários, S.A.”

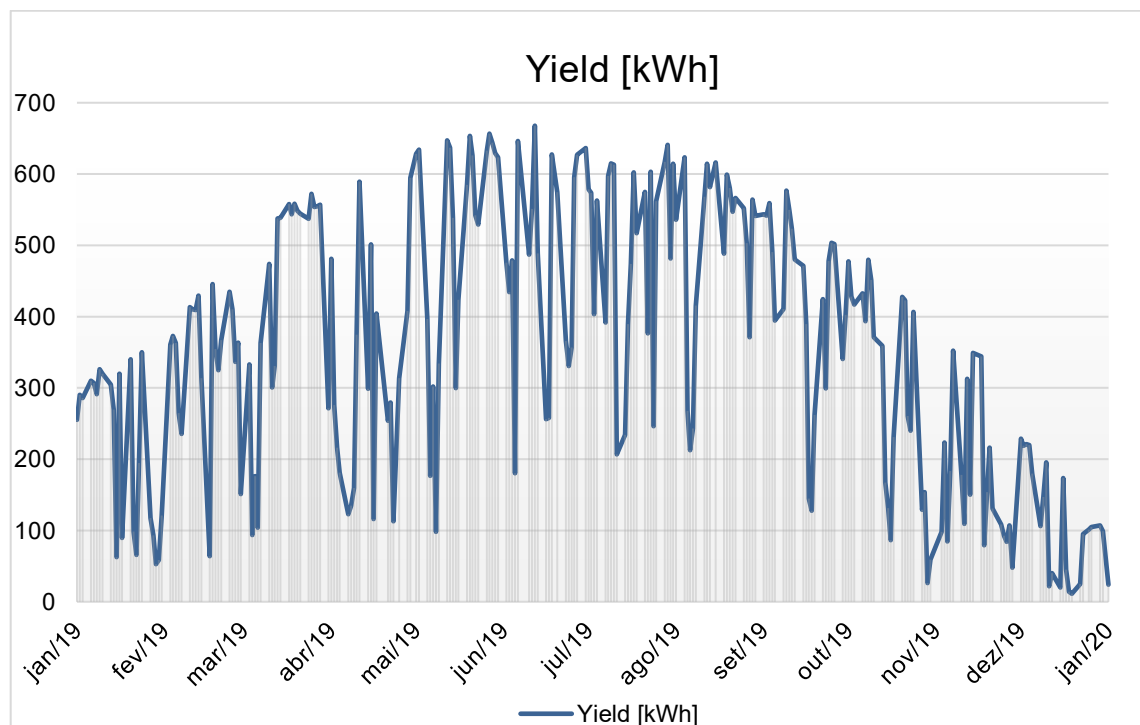
A segunda aplicabilidade de análise preditiva na “OLI – Sistemas Sanitários, S.A.” foi implementada na produção do painel solar fotovoltaico em Killowatt, pertencente à empresa “OLI Moldes”, operacional desde janeiro de 2019.

Portanto, o objetivo do caso de estudo é determinar os próximos valores do consumo de energia.

Primeiramente, foi recolhida a informação da produção de energia do painel solar referente a 1 ano, isto é, de 2 de janeiro de 2019 a 2 de janeiro de 2020. Uma vez que os painéis solares não estão operacionais aos fins-de-semana, feriados e outras folgas (nomeadamente, 1 de janeiro, 19 e 25 de abril, 1 de maio, 10 e 20 de junho, 15 de agosto, 1 de novembro e 25 de dezembro), foram retirados dos dados, para evitar o enviesamento da análise.

Os dados em questão estão representados na figura 10.

Figura 10  
Dados do Consumo de Energia do Painel Solar



Em prol de complementar a análise, foi reunida informação pertinente, para depois, utilizá-la para testar a precisão da previsão de algoritmos, e assim, por fim, adotar o método que mais se adequa ao propósito.

Para a constituição eficaz e eficiente deste projeto, foi, paralelamente, recolhida, a informação meteorológica passada.

A temperatura em Fahrenheit tipicamente a 2 metros (tmpf), a temperatura Dew Point em Fahrenheit tipicamente a 2 metros (dwpf), a humidade relativa (relh), a direção do vento em graus do norte (drct), a velocidade do vento em Knots (sknt), a precipitação de uma hora do período entre o tempo de observação e o horário de redefinição anterior à precipitação, em polegadas (p01i) (o valor pode ou não conter precipitação congelada derretida por um dispositivo do sensor ou ser estimado por outros meios, não existindo um banco de dados autorizados que denote o sensor de cada localidade), o altímetro de pressão em polegadas (alti), a pressão do nível do mar em milibar (mslp), a visibilidade em milhas (vsby), a rajada de vento em Knots (gust), a

cobertura dos níveis 1 a 4 do céu (skc1, skc2, skc3 e skc4) e a altitude dos níveis 1 a 4 do céu em “pés” (sky11, sky12, sky13 e sky14), os códigos climáticos atuais (presentwx), observação relatada não processada no formato METAR<sup>1</sup> (metar) foram os indicadores veiculados pelas bibliotecas Curl e Riem. Também disponibilizam a criação de gelo, os picos da direção do vento, da rajada de vento e do tempo do vento, e a sensação, em que apenas a última variável apresentou dados, embora não conclusivos.

O código posterior indica a recolha de variáveis meteorológicas de Ovar. Foram utilizadas as bibliotecas Riem e Curl, em que a função `riem_networks` revela a lista dos códigos dos países disponíveis para a obtenção de dados. Neste caso, o país que se pretende é Portugal, em que o código é “PT\_\_ASOS”. A função `riem_stations` funciona de forma idêntica, com a diferença de que diz respeito a localidades de determinado país. Assim, foi inserida a determinação de Portugal para obter os códigos da localidade de interesse (“LPOV” corresponde a Ovar).

A função `riem_measures` viabiliza a demonstração das variáveis meteorológicas da localidade pretendida, com os limites temporais desejáveis.

A visualização no R da tabela do objeto é alcançada através do código “View”. Note-se que, ao contrário de maior parte das funções do R descritas neste trabalho, a letra maiúscula é exigida para o funcionamento da mesma.

Através de “summary”, são indicados o valor mínimo, o 1º e o 3º Quartis, a média, a mediana, a média, o valor máximo, os dados não observáveis, o tamanho, o modo e a classe.

---

<sup>1</sup> O formato METAR é um formato que serve para divulgar informação climática, como por exemplo a temperatura.



```
# Clima Histórico de Ovar (estação mais próxima de Aveiro)

library(riem)
library(curl)

DT <- riem_networks()
View(DT)
PT <- riem_stations("PT__ASOS")
View(PT)
OV <- riem_measures("LPOV", date_start = "2019-01-02", date_end = "2020-01-02")
View(OV)
summary(OV)
```

Como a temperatura e o Dew Point estão em Fahrenheit, foi necessário a passagem para graus Celsius, do seguinte modo:

```
# 1. FtoC
#T(°C) = (T(°F) - 32) × 5/9 ou
#T(°C) = (T(°F) - 32) / (9/5) ou
#T(°C) = (T(°F) - 32) / 1.8

OV$tmpc=(OV$tmpf-32)*5/9
OV$tmpc
OV$dwpc=(OV$dwf-32)*5/9
OV$dwpc

# 2. FtoC

library(weathermetrics)

fahrenheit.to.celsius(T.fahrenheit = 53)
```

O ponto 1 e o ponto 2 revelam formas alternativas de realizar a operação.

O mesmo foi feito para a velocidade do vento, em que se passou de Knots para Quilómetros (veja-se o código em baixo).

```
# Forma de converter Knot em Km
##'wind speed'

OV$skph=OV$sknt*1.852
OV$skph
```

Assim, a temperatura em graus Celsius, o Dew Point em graus Celsius e a velocidade do vento em quilómetros é representado por tmpc, dwpc e skph, respetivamente.

Para guardar os dados num Excel e transferir o mesmo para o R, foi escrito o seguinte código:

```
write.xlsx (x = OV , file = "OLI - Clima de Ovar.xlsx", sheet = "Dados_Brutos_OV",
           col.names = TRUE, row.names = TRUE, append = FALSE)

clima <- read.xlsx("OLI - Informações Climatéricas - 2019.xlsx", sheet = "Sheet")
```

Com o mesmo objetivo, foi realizado um código semelhante, para agregar a informação da produção solar (dado o nome de solarlog) com as informações climáticas (dado o nome de clima). Assim, a função `cbind` serviu para a combinação de observações, e a função `write.xlsx` para escrever no Excel essa informação.

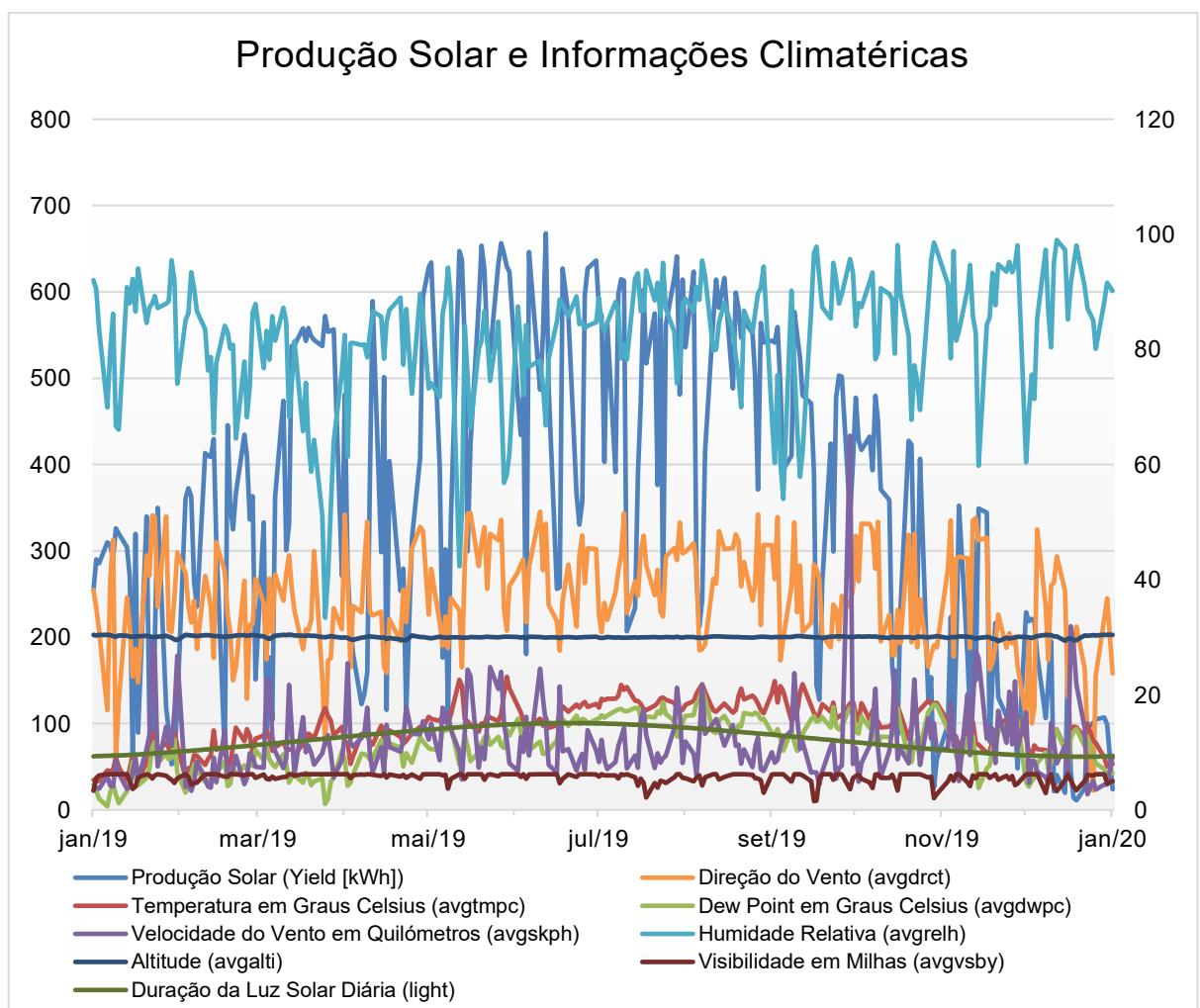
```
cbind(solarlog, clima)
solima <- cbind(solarlog, clima)

write.xlsx (x = solima , file = "OLI - solima.xlsx", sheet = "Sheet1",
           col.names = TRUE, row.names = TRUE, append = FALSE)
```

Um gráfico que pode descrever estes dados é o gráfico da figura 11.

Figura 11

Produção da Energia e as Informações Climatéricas



A biblioteca Suncalc revela a indicação das horas em que se inicia o nascer do dia e o pôr-do-sol, na localidade pretendida através da latitude e longitude. A função `seq.date` orienta os limites temporais e o espaçamento entre os mesmos (no exemplo, está indicado que é por 1 dia em “by = 1”). Esta função tem uma desvantagem, uma vez que é necessário alterar os números em `Sys.Date` regularmente. O -380 e o -14 no código em baixo estará em constante mudança, pois o -380 indica que o limite mínimo temporal é o de há 380 dias. Assim, no dia a seguir, terá de ser alterado. O mesmo se denota com o -14, que é o limite temporal máximo.

```
# Nascer do dia e pôr-do-sol
library(suncalc)

sunlight <- getsunlightTimes(date = seq.Date(Sys.Date()-380, Sys.Date()-14, by =1),
                             lat = 40.64427, lon = -8.64554, data = NULL,
                             keep = c("sunrise","sunset"), tz = "GMT")

sunlight
```

Para a obtenção do intervalo de tempo entre o nascer do dia e do pôr-do-sol, foi criado um parâmetro, com a medida de horas, para assim determinar a sua correlação com a produção dos painéis fotovoltaicos, bem como as outras variáveis anteriormente descritas.

```
# Luz solar por Dia (diferença entre o pôr-do-sol e o nascer do dia)

s.set <- sunlight$sunset
s.rise <- sunlight$sunrise
light <- s.set - s.rise

cbind (sunlight, light)
suncalc <- cbind (sunlight, light)
plot(light)
```

Mais uma vez, é utilizada a função `cbind` para juntar os dados do nascer do dia, o pôr do sol, e a diferença entre as mesmas. O `plot` designa-se para relatar graficamente os dados.

A combinação foi transferida para o Excel, como já atrás realizado, e agregado toda a informação pertinente numa tabela, designada de “all”.

```
write.xlsx(suncalc, file = "suncalc.xlsx", sheetName="Sheet1",
          col.names=TRUE, row.names=TRUE, append=FALSE)

sunc <- read.xlsx("OLI - Suncalc.xlsx", sheet = "suncalc")

sunsol <- cbind(sunc, solarlog)

all <- cbind(sunsol, clima)
```

Assim, com o cruzamento destas informações, foram determinadas as correlações das variáveis temperatura em graus Celcius (tmpc), o Dew Point em graus Celsius (dwpc), a altitude de pressão em polegadas, a visibilidade em milhas (vsby), a direção do vento em graus do Norte (drct), a humidade relativa (relh), a velocidade do vento em quilómetros (skph) e a diferença entre o pôr do sol e o nascer do dia (light), face à produção de energia do painel solar. Uma nota importante a destacar é a exclusão das restantes variáveis mencionadas anteriormente, pelos seus dados insuficientes, e assim, tenderem a uma análise pouca conclusiva e enviesada. A imagem em baixo descreve o código para tal efeito.

```
# Correlações
library(corr)
library(corrplot)
library(RColorBrewer)

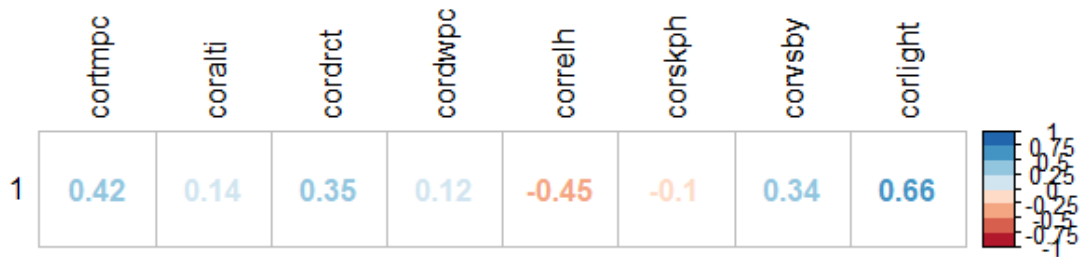
cortmpc <- cor(solima$`yield [kwh]`, solima$avgtmpc)
coralti <- cor(solima$`yield [kwh]`, solima$avgalti)
cordrct <- cor(solima$`yield [kwh]`, solima$avgdrct)
cordwpc <- cor(solima$`yield [kwh]`, solima$avgdwpc)
correlh <- cor(solima$`yield [kwh]`, solima$avgrelh)
corskph <- cor(solima$`yield [kwh]`, solima$avgskph)
corvsby <- cor(solima$`yield [kwh]`, solima$avgvsby)
corlight <- cor(solima$`yield [kwh]`, sunsol$light)

gcor <- cbind(cortmpc, coralti, cordrct, cordwpc, correlh, corskph, corvsby, corlight)
corrplot(gcor, method = "number", type="upper", tl.col="black", col=brewer.pal(n=8, name="RdBu"))
```

Assim, o código apresentado em cima permite produzir correlações entre a produção de energia e cada uma das variáveis que, potencialmente, pode explicar essa produção. O seu output é mostrado na Figura 12.

Figura 12

Correlações das Variáveis Climáticas com a Produção de Energia



A correlação pode ir do -1 a 1. Quando a correlação entre duas variáveis é 1, estamos perante uma correlação positiva perfeita, enquanto o valor -1 indica uma correlação negativa perfeita. Um valor baixo, tanto negativo como positivo, indica uma fraca correlação. Assim, por exemplo, o Dew Point em graus Celsius (dwpc) tem uma fraca relação positiva com a produção de energia, enquanto que a velocidade do vento em quilómetros (skph) tem uma correlação negativa fraca.

Pode-se verificar que as correlações positivas mais altas são a duração da luz solar (light), a temperatura em graus Celsius (tmpc), a direção do vento em graus do Norte (drct), por ordem decrescente. A humidade relativa tem uma correlação negativa significativa, o que quer dizer que quando uma variável aumenta, a outra decresce.

Foram utilizados o modelo de regressão linear simples, o modelo de regressão linear múltipla, o modelo de regressão Lasso, o modelo de regressão de Ridge e o modelo de floresta aleatória.

Para os modelos de regressão lineares foram tidas em conta as oito variáveis já mencionadas: a temperatura em graus Celsius (tmpc), a direção do vento em graus do Norte (drct), a visibilidade em milhas (vsby), a humidade relativa (relh), a duração da luz solar (light), a velocidade do vento em quilómetros (skph), o Dew Point em graus Celsius (dwpc) e a altitude (alti).

O modelo de regressão linear foi aplicado a todas as variáveis em questão (considerando uma a uma), no qual se pode verificar nos códigos em baixo.

```
# Modelo de Regressão Linear Simples

mydata <- read_xlsx("OLI - Dados Combinados 02.01.19 a 02.02.20.xlsx", sheet = "Folha1")

model1 <- lm(`Yield [kwh]`~ avgtmpc, data=mydata)
summary(model1)
accuracy(model1)

model2 <- lm(`Yield [kwh]`~ avgskph, data=mydata)
summary(model2)
accuracy(model2)

model3 <- lm(`Yield [kwh]`~ light, data=mydata)
summary(model3)
accuracy(model3)

model4 <- lm(`Yield [kwh]`~ avgrelh, data=mydata)
summary(model4)
accuracy(model4)

model5 <- lm(`Yield [kwh]`~ avgvsby, data=mydata)
summary(model5)
accuracy(model5)

model6 <- lm(`Yield [kwh]`~ avgdrct, data=mydata)
summary(model6)
accuracy(model6)

model7 <- lm(`Yield [kwh]`~ avgalti, data=mydata)
summary(model7)
accuracy(model7)

model8 <- lm(`Yield [kwh]`~ avgdwpc, data=mydata)
summary(model8)
accuracy(model8)
```

Os melhores resultados foram obtidos no modelo construído pela produção de energia solar e pela luz solar, em que o  $R^2$  e o  $R$  ajustado é de, aproximadamente, 44%, (como podemos ver em baixo), e o  $p - value$  é inferior às demais.

```
> summary(model3)

Call:
lm(formula = `Yield [kwh]` ~ light, data = mydata)

Residuals:
    Min       1Q   Median       3Q      Max
-384.92 -115.46   34.03  121.52  211.18

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -396.824     54.490   -7.283 4.22e-12 ***
light           61.850       4.416  14.005 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 140.3 on 251 degrees of freedom
Multiple R-squared:  0.4386,    Adjusted R-squared:  0.4364
F-statistic: 196.1 on 1 and 251 DF,  p-value: < 2.2e-16

> accuracy(model3)
```

	ME	RMSE	MAE	MPE	MAPE	MASE
Training set	-1.724411e-15	139.7745	121.127	-53.30315	79.09624	0.754468

Também nas medidas de precisão, nomeadamente o ME, RMSE, MAE, MPE, MAPE e MASE, apresentaram-se, em quase todas elas, inferiores comparativamente aos restantes modelos (conferindo-se em baixo).

```
> accuracy(model1)
      ME      RMSE      MAE      MPE      MAPE      MASE
Training set -3.870723e-14 169.3542 145.961 -89.35451 118.2651 0.9091525
> accuracy(model2)
      ME      RMSE      MAE      MPE      MAPE      MASE
Training set 4.336452e-15 185.5467 159.6337 -96.7372 125.5306 0.9943156
> accuracy(model3)
      ME      RMSE      MAE      MPE      MAPE      MASE
Training set -1.724411e-15 139.7745 121.127 -53.30315 79.09624 0.754468
> accuracy(model4)
      ME      RMSE      MAE      MPE      MAPE      MASE
Training set 2.524221e-15 166.2258 143.1244 -73.7259 100.5103 0.8914838
> accuracy(model5)
      ME      RMSE      MAE      MPE      MAPE      MASE
Training set -7.543922e-15 175.3064 151.6276 -76.03537 104.2571 0.944448
> accuracy(model6)
      ME      RMSE      MAE      MPE      MAPE      MASE
Training set 2.007172e-14 175.075 148.8194 -88.34263 115.953 0.9269563
> accuracy(model7)
      ME      RMSE      MAE      MPE      MAPE      MASE
Training set 1.159174e-14 184.6868 158.6327 -94.21703 122.7455 0.988081
> accuracy(model8)
      ME      RMSE      MAE      MPE      MAPE      MASE
Training set 1.792506e-14 185.2 157.8707 -100.8528 129.5284 0.9833343
```

O modelo de regressão linear múltiplo, com todas variáveis, foi obtido através do código em baixo.

```
# Modelo de Regressão Linear Múltipla
mydata <- read_xlsx("OLI - Dados Combinados 02.01.19 a 02.02.20.xlsx", sheet = "Folha1")
modelmulti <- lm(`Yield [kwh]`~ avgtmpc + avgdrct + avgrelh + light + avgvsby
+ avgdwpc + avgalti + avgskph, data=mydata)
summary(modelmulti)
accuracy(modelmulti)
```

No modelo verifica-se que o  $R^2$  é 67% e o R ajustado é de 66%, com um  $p$  – *value* bastante próximo de 0.

Analisando cada variável, verificamos que três variáveis – a direção do vento em graus do Norte (drct), a altitude (alti) e a duração de luz solar (light) se destacam pelo seu  $p$  – *value* bastante pequeno. Pode-se inferir tal afirmação através dos resultados em baixo.

```
> summary(modelmulti)

Call:
lm(formula = `Yield [kwh]` ~ avgtmpc + avgdrct + avgrelh + light +
  avgvsby + avgdwpc + avgalti + avgskph, data = mydata)

Residuals:
    Min       1Q   Median       3Q      Max
-303.08  -72.69   10.18   81.65  263.38

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -7218.3765   1404.2607  -5.140 5.62e-07 ***
avgtmpc       43.2866    22.4202    1.931  0.0547 .
avgdrct       0.6976     0.1312    5.316 2.40e-07 ***
avgrelh       4.5501     5.6661    0.803  0.4227
light       51.9565     5.2408    9.914 < 2e-16 ***
avgvsby      19.3881    10.5621    1.836  0.0676 .
avgdwpc     -41.5652    23.7298   -1.752  0.0811 .
avgalti     204.9866    40.8646    5.016 1.02e-06 ***
avgskph      -3.1644     1.2304   -2.572  0.0107 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 109 on 244 degrees of freedom
Multiple R-squared:  0.671,    Adjusted R-squared:  0.6602
F-statistic: 62.21 on 8 and 244 DF,  p-value: < 2.2e-16

> accuracy(modelmulti)

              ME          RMSE          MAE          MPE          MAPE          MASE
Training set 1.733352e-15 107.0035 88.48791 -19.30499 51.28239 0.5511677
```

As mesmas variáveis foram inseridas nos modelos de floresta aleatória e no Lasso, isto é, a temperatura em graus Celsius (tmpc), a direção do vento em graus do Norte (drct), a visibilidade em milhas (vsby), a humidade relativa (relh), a duração da luz solar (light), a velocidade do vento em quilómetros (skph), o Dew Point em graus Celsius (dwpc) e a altitude (alti).

Com as bibliotecas Caret, randomForest e caTools é realizado o modelo de floresta aleatória. No modelo de floresta aleatória inseriu-se os dados com as variáveis, como já ilustrado em casos anteriores e no código abaixo.

```
### Random Forest ###
mydata <- read_xlsx("OLI - Dados Combinados 02.01.19 a 02.02.20.xlsx", sheet = "Folha1")
Rf_fit<-randomForest(mydata$`Yield [kwh]` ~ mydata$avgtmpc + mydata$avgdwpc" +
  mydata$`avgskph` + mydata$`avgrelh` + mydata$`avgdrct` +
  mydata$`avgvsby` + mydata$`light` + mydata$`avgalti`, data = mydata)
```

De seguida, para se realizar a previsão, é necessário a divisão dos dados já observáveis, uma vez que se pretende verificar os supostos resultados num determinado espaço temporal, em comparação com o real.



```
# Splitting the data into test and train

sample = sample.split(mydata$`Yield [kwh]`, splitRatio = .5)
train = subset(mydata, sample == TRUE)
test = subset(mydata, sample == FALSE)
```

Para obter a previsão, é utilizada a função predict, que irá prever nos novos dados temporais acima divididos.

```
# Results

rf <- randomForest( `Yield [kwh]` ~ avgtmpc + avgdwpc + avgskph + avgrelh + avgdrct +
                    avgvsby + light + avgalti, data=train)
pred = predict(rf, newdata=test)
pred
head(pred)

cbind(test, pred)
pred_random <- cbind(test, pred)

write.xlsx(pred_random, file = "OLI - Resultado Random Forest.xlsx", sheetName = "Folha1")
```

A função head revela uma porção das previsões realizadas:

```
> head(pred)
      1      2      3      4      5      6
238.2962 272.9322 311.6788 313.5148 317.0111 326.2365
```

Por fim, verifica-se a previsão do modelo em questão, com o seguinte código:

```
# Accuracy

postResample(pred_random$pred, pred_random$`Yield [kwh]`)
print(Rf_fit)
```

O mesmo deu origem aos resultados representados abaixo.

```
> postResample(pred_random$pred, pred_random$`Yield [kwh]`)
      RMSE      Rsquared      MAE
110.8305532  0.6755186  92.2841573
> print(Rf_fit)

call:
randomForest(formula = mydata$"Yield [kwh]" ~ mydata$avgtmpc + mydata$avgdwpc + mydata$avgskph + mydata$avgrelh + mydata$avgdrct + mydata$avgvsby + mydata$light + mydata$avgalti, data = mydata)
Type of random forest: regression
Number of trees: 500
No. of variables tried at each split: 2

Mean of squared residuals: 10533.48
% var explained: 69.73
```

É apresentado (em baixo) uma parte dos resultados obtidos por este modelo, onde o [1] equivale ao período, o [2] à produção de energia e o [3] à previsão.

	[,1]	[,2]	[,3]
[1,]	2	290.19	238.2962
[2,]	3	285.80	272.9322
[3,]	4	309.80	311.6788
[4,]	6	291.16	313.5148
[5,]	7	325.97	317.0111
[6,]	8	320.06	326.2365

O modelo de lasso no R inicia-se de forma semelhante aos restantes, isto é, a inserir os dados.

```
### Lasso Regression ###
# Loading the data
mydata <- read_xlsx("OLI - Dados Combinados 02.01.19 a 02.02.20.xlsx", sheet = "Folha1")
x_vars <- model.matrix(mydata$Yield [kwh] ~ mydata$avgtnpc + mydata$avgdwpc + mydata$avgskph +
  mydata$avgrelh + mydata$avgdrct + mydata$avgvsby + mydata$light + mydata$avgaltf,
  data = mydata) [, -1]
y_var <- mydata$Yield [kwh]
lambda_seq <- 10^seq(2, -2, by = -.1)
```

Segue a mesma lógica do modelo de floresta aleatória para realizar o teste/treino dos dados.

```
# Splitting the data into test and train

set.seed(86)
train = sample(1:nrow(x_vars)/2)
test <- setdiff(1:nrow(x_vars), train)
x_test = (-train)
y_test = y_var[test]

cv_output <- cv.glmnet(x_vars[train,], y_var[train],
  alpha = 1, lambda = lambda_seq)
cv_output
```

```
> cv_output

Call: cv.glmnet(x = x_vars[train, ], y = y_var[train], lambda = lambda_seq, alpha = 1)

Measure: Mean-Squared Error

      Lambda Measure   SE Nonzero
min  0.158   11233   972         7
1se 10.000   12006  1021         6
```

O resultado em cima descrito irá servir para encontrar o melhor  $\lambda$ , o parâmetro de ajustamento de Lasso. Assim foi configurada a junção entre o cv\_output e o valor mínimo de  $\lambda$ , a qual gerou o número 0,1584893 (dado o nome de bestlam no R).

De seguida, foi otimizado o modelo, com o melhor  $\lambda$  e foram feitas as previsões (tal como no modelo anterior, são feitas com os dados de treino).

```
# Rebuilding the model with best lamda value identified
lasso_best <- glmnet(x_vars[train,], y_var[train], alpha = 1, lambda = best_lam)
lasso_best
pred <- predict(lasso_best, s = best_lam, newx = x_vars[test,])
pred
final <- cbind(y_var[test], pred)
view(final)
head(final)

write.xlsx(final, file = "OLI - Resultado Preditivo Lasso.xlsx", sheetName = "Folha1")
```

Uma porção dos resultados é demonstrado em baixo.

			1
127	403.21	480.8157	
128	563.02	493.1218	
129	391.50	459.5215	
130	597.30	460.0782	
131	614.84	502.2143	
132	613.02	493.5833	

Por fim, com a função print, são demonstrados os erro médio e o erro médio absoluto e o  $R^2$ .

```
> postResample(actual, preds)
      RMSE      Rsquared      MAE
212.9588403  0.0439837 177.5851044
```

Para o problema em questão, pode-se concluir que os erros médios são elevados e o  $R^2$  é baixo (4.4%).

O modelo de regressão de Ridge foi executado da mesma forma que foi realizado o modelo de regressão de Lasso.

Realizou-se o teste/treino, em que se foi sugerido o seguinte resultado:

```
> cv_output

Call: cv.glmnet(x = x_vars[train, ], y = y_var[train], lambda = lambda_seq, alpha = 0)

Measure: Mean-Squared Error

      Lambda Measure      SE Nonzero
min  0.25  11242  971.4      8
1se 39.81  12102 1020.5      8
```

O melhor resultado para o  $\lambda$  foi de 0,2511886. Assim, foi otimizado o modelo, com o  $\lambda$  sugerido pelo R, e realizadas as previsões. Assim, pode-se verificar um excerto dos resultados obtidos.

```

1
127 403.21 481.1397
128 563.02 493.3828
129 391.50 459.7651
130 597.30 460.2804
131 614.84 502.8071
132 613.02 494.0915

```

A tabela (9) ilustra uma comparação dos métodos utilizados dos dados reais da produção de energia face aos resultados previstos, num período temporal de 20 dias (onde os “Model” de 1 a 8, definidos anteriormente, representam os modelos de regressão linear simples com cada uma das variáveis, e o “Model Multi”, representa o modelo de regressão linear múltiplo).

Tabela 9

Comparação do Resultados Preditivos dos Modelos Causais (Modelo de Regressão Linear Simples, Modelo de Regressão Linear Múltiplo, Modelo de Floresta Aleatória, Modelo de Regressão de Lasso e Modelo de Regressão de Ridge).

Data	Dados Reais	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model Multi	F. Aleatória	Lasso	Ridge
05/07/2019	563,02	447,90	369,99	530,52	337,32	380,64	331,75	351,26	351,26	497,93	490,84	493,12	493,59
11/07/2019	613,02	477,91	349,01	524,50	365,01	386,23	463,01	338,25	338,25	555,54	490,67	493,58	493,97
12/07/2019	206,86	487,92	369,58	523,30	385,97	397,41	340,09	336,89	336,89	509,94	524,04	498,21	499,30
16/07/2019	389,62	441,08	369,40	517,99	271,46	249,85	361,47	340,65	340,65	423,13	348,08	406,90	406,96
17/07/2019	473,82	431,10	344,59	516,55	321,68	344,16	416,09	346,33	346,33	484,18	548,44	427,10	427,04
22/07/2019	575,08	424,45	369,56	508,60	307,73	286,40	362,34	352,05	352,05	456,03	473,88	450,84	451,11
23/07/2019	376,25	434,11	361,87	506,88	283,96	228,71	393,50	348,89	348,89	437,16	414,86	411,77	411,80
24/07/2019	603,12	449,52	366,83	505,12	327,82	316,82	342,36	344,33	344,33	447,91	505,16	434,53	435,09
25/07/2019	246,43	452,77	364,57	503,32	257,57	305,61	336,22	343,82	343,82	403,83	323,47	373,95	373,99
02/08/2019	535,92	428,07	356,04	487,60	304,90	346,55	415,33	351,26	351,26	469,98	482,51	428,86	428,99
09/08/2019	415,53	472,30	352,10	472,19	275,36	356,75	301,43	348,12	348,12	370,97	255,55	320,57	320,95
12/08/2019	568,33	408,61	346,69	465,21	373,81	397,41	382,46	373,00	373,00	498,86	566,72	463,64	463,95
19/08/2019	488,23	401,61	347,66	448,21	348,37	397,41	420,19	354,52	354,52	470,33	564,15	419,89	419,99
02/09/2019	543,62	467,36	361,41	411,99	462,08	393,69	423,52	343,82	343,82	536,95	515,22	499,47	499,90

Como nos outros modelos demonstrados, foi elaborada uma comparação das medidas de precisão (ver tabela 10), neste caso, o RMSE, o  $R^2$  e o MAE, medidas já descritas no capítulo 3.3.

Tabela 10

*Comparação das Medidas de Precisão dos Modelos Causais (Modelo de Regressão Linear Simples, Modelo de Regressão Linear Múltiplo, Modelo de Floresta Aleatória, Modelo de Regressão de Lasso e Modelo de Regressão de Ridge).*

<b>Modelos</b>	<b>RMSE</b>	<b>R<sup>2</sup></b>	<b>MAE</b>
Model 1	169,35	0,18	145,96
Model 2	185,55	0,01	159,63
Model 3	139,77	0,44	121,13
Model 4	166,23	0,21	143,12
Model 5	175,31	0,12	151,63
Model 6	175,01	0,12	148,82
Model 7	184,69	0,02	158,63
Model 8	185,2	0,01	157,87
Model Multi	107,01	0,67	88,49
Modelo de Ridge	130,57	62,82%	107,11
Modelo de Lasso	130,52	62,87%	107,18
Modelo de Floresta Aleatória	114,71	0,61	93,73

Verifica-se que, efetivamente, o Modelo de Regressão Linear Múltiplo apresenta os melhores resultados nas medidas de precisão, sendo que o pior é o modelo com a variável da velocidade do vento em quilómetros (skph).

Assim, podemos afirmar que o melhor método a ser utilizado será o Modelo de Regressão Linear Múltiplo.

# Capítulo 5

## 5. Conclusões

A análise preditiva é um ativo bastante importante para a funcionalidade eficaz e eficiente das organizações. Através da mesma, as instituições empresariais podem planejar ações estratégicas e tomar decisões de forma mais adequada.

Este trabalho aplicou vários métodos de previsão para obter conhecimento empírico sobre qual deles seria mais adequado, tanto para prever as vendas totais da empresa, como também para prever o consumo energético do painel solar, dentro do período temporal estipulado. Este conhecimento teria por base tanto os resultados estimados, como as métricas de precisão descritas.

Assim, relativamente ao primeiro caso de estudo, foram tidos em conta métodos quantitativos de séries temporais, nomeadamente os métodos Holt Winters aditivo e multiplicativo e o método ARIMA, e calculados o ME, RMSE, MAE, MAPE, MASE e ACF.

Ao verificar os seus resultados preditivos e de precisão, constata-se que o método ARIMA se assemelha mais com a realidade, com medidas de erro inferiores, aquando a sua comparação com os demais.

No âmbito do segundo caso, teve-se em conta os métodos quantitativos de regressão linear, de regressão múltipla, de regressão de Lasso, de regressão de Ridge e de Floresta Aleatória.

Para este caso, os procedimentos foram os mesmos comparativamente ao anterior, isto é, analisou-se as suas respetivas previsões e as medidas de precisão.

No segundo caso analisado, verificamos que o Modelo de Regressão Linear Múltiplo desempenha a melhor performance, comparativamente aos restantes modelos.

# Bibliografia

- Boonsiritomachai, W., McGrath, G. M., & Burgess, S. (2016). Exploring business intelligence and its depth of maturity in Thai SMEs. *Cogent Business and Management*, 3(1). <https://doi.org/10.1080/23311975.2016.1220663>
- Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2016). *Time Series Analysis: Forecasting and Control*. John Wiley & Sons, Inc.
- Brockwell, P. J., & Davis, R. A. (1990). *Time Series: Theory and Methods. Journal of the Royal Statistical Society. Series A (Statistics in Society)* (Vol. 153). <https://doi.org/10.2307/2982983>
- Caiado, J. (2016). *Métodos de Previsão em Gestão - Com Aplicações em Excel* (2nd Edition). Edições Sílabo.
- Chatfield, C. (2000). *Time-Series Forecasting. Urologiia (Moscow, Russia : 1999)*. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/16856455>
- Chen, H., H.L.Chiang, R., & C. Storey, V. (2012). Business Intelligence and Analytics: From Big Data To Big Impact. *MIS Quarterly*, 36(4), 1165–1188. [https://doi.org/10.1016/S0140-6736\(09\)61833-X](https://doi.org/10.1016/S0140-6736(09)61833-X)
- Contreras, J., Espínola, R., Nogales, F. J., & Conejo, A. J. (2003). ARIMA models to predict next-day electricity prices. *IEEE Transactions on Power Systems*, 18(3), 1014–1020. <https://doi.org/10.1109/TPWRS.2002.804943>
- Cowpertwait, P. S. P., & Metcalfe, A. V. (2009). *Introductory Time Series with R*. Springer. <https://doi.org/10.1007/978-0-387-88698-5>
- Few, S. (2005). Dashboard Design : Beyond Meters , Gauges , and Traffic Lights, 18–24.
- Hans, C. (2009). Bayesian lasso regression. *Biometrika*, 96(4), 835–845. <https://doi.org/10.1093/biomet/asp047>
- Harrell, F. E. (2015). *Regression modeling strategies: with applications to linear*



- models, logistic regression, and survival analysis. Statistical Methods in Medical Research* (2nd Editio, Vol. 13). Springer.  
<https://doi.org/10.1177/096228020401300512>
- Hartmann, P. M., Zaki, M., & Feldmann, N. (2014). Big Data for Business? A Taxonomy of Data-driven Business Models used by Start-up Firms.
- Hillmer, S. C., & Wei, W. W. S. (1991). *Time Series Analysis: Univariate and Multivariate Methods. Journal of the American Statistical Association* (Vol. 86).  
<https://doi.org/10.2307/2289741>
- Hipel, K. W., & McLeod, I. A. (1994). *Time Series Modelling of Water Resources and Environmental Systems*. Elsevier.
- Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice. Principles of Optimal Design*. <https://doi.org/10.1017/9781316451038.010>
- Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4), 679–688.  
<https://doi.org/10.1016/j.ijforecast.2006.03.001>
- Hyndman, R. J., Koehler, A. B., Ord, J. K., & Snyder, R. D. (2008). *Forecasting with Exponential Smoothing: The State Space Approach*. Springer.  
<https://doi.org/10.1007/978-3-540-71918-2>
- Hyndman, R. J., Wheelwright, S. C., & Makridakis, S. (1998). *Forecasting, Methods and Applications*.
- Ihaka, R., & Gentleman, R. (1996). R: A Language for Data Analysis and Graphics. *Computational and Graphical Statistics*, 5(3), 299–314.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: With Applications in R. Springer Texts in Statistics*. Springer. <https://doi.org/10.1016/j.peva.2007.06.006>
- Keim, D. A. (2002). Information Visualization and Visual Data Mining. *IEEE Transactions on Visualization and Computer Graphics*, 7(1), 100–107.  
<https://doi.org/10.1109/2945.981847>

- Lee, P. M. (2013). Use Of Data Mining In Business Analytics To Support Business Competitiveness. *Review of Business Information Systems (RBIS)*, 17(2), 53–58. <https://doi.org/10.19030/rbis.v17i2.7843>
- Lim, E. P., Chen, H., & Chen, G. (2013). Business Intelligence and Analytics : Research Directions, 3, 1–10. <https://doi.org/https://doi.org/10.1145/2407740.2407741>
- Lustig I, Dietrich B, Johnson C, Dziekan C. (2010). The analytics journey. Analytics.
- Manyika, J., Chui, M., Brown, B., Bughin, J., & Dobbs, R. (2011). Big data: The Next Frontier for Innovation, Competition, and Productivity. *McKinsey Global Institute*, 156. <https://doi.org/10.1080/01443610903114527>
- Olszak M., C., & Ziemba, E. (2007). Approach to Building and Implementing Business Intelligence Systems. *Journal of Information, Knowledge and Management*, 2, 135–148.
- Organization for Economic Co-operation and Development. (2013). Exploring data-driven innovation as a new source of growth. *OECD Digital Economy Papers*, (April). <https://doi.org/10.1787/5k437p2rp4bq-en>
- Ranjan, J. (2009). Business Intelligence: Concepts, Components, Techniques and Benefits. *Journal of Theoretical and Applied Information Technology*, 9(1), 60–70. <https://doi.org/10.2139/ssrn.2150581>
- Rubin, V. L., & Lukoianova, T. (2013). Veracity Roadmap: Is Big Data Objective, Truthful and Credible? *Advances in Classification Research Online*, 24(1), 4–15. <https://doi.org/10.7152/acro.v24i1.14671>
- Schläfke, M., Silvi, R., & Möller, K. (2013). A Framework for Business Analytics in Performance Management. *International Journal of Productivity and Performance Management*, 62(1), 110–122. <https://doi.org/10.1108/17410401311285327>
- Shanks, G., Cosic, R., & Maynard, S. (2012). Towards a Business Analytics

Capability Maturity Model, 1–11.

Slonneger, K., & Kurtz, B. L. (1995). *Formal Syntax and Semantics of Programming Languages* (Vol. 340). <https://doi.org/10.1017/CBO9781107415324.004>

Waller, M. A., & Fawcett, S. E. (2013). Data Science, Predictive Analytics, and Big Data: A Revolution that will transform Supply Chain Design and Management. *Journal of Business Logistics*, 34(2), 77–84. <https://doi.org/10.1111/jbl.12010>

Yang, H. H., & Moody, J. (1999). Data Visualization and Feature Selection: New Algorithms for Nongaussian Data. *Advances in Neural Information Processing Systems*, 12, 687–693. Retrieved from <https://papers.nips.cc/paper/1779-data-visualization-and-feature-selection-new-algorithms-for-nongaussian-data.pdf>  
<http://papers.nips.cc/paper/1779-data-visualization-and-feature-selection-new-algorithms-for-nongaussian-data.pdf>